2019

Real-time Sound Source Separation For Music Applications

Dan Barry
Technological University Dublin

Real-time Sound Source Separation For Music Applications

Dan Barry (MPhil)

Submitted for the award of PhD

Technological University Dublin

School of Electrical and Electronic Engineering

Supervisor: Dr. David Dorran

2019

ABSTRACT

Sound source separation refers to the task of extracting individual sound sources from some number of mixtures of those sound sources. In this thesis, a novel sound source separation algorithm for musical applications is presented. It leverages the fact that the vast majority of commercially recorded music since the 1950s has been mixed down for two channel reproduction, more commonly known as stereo. The algorithm presented in Chapter 3 in this thesis requires no prior knowledge or learning and performs the task of separation based purely on azimuth discrimination within the stereo field. The algorithm exploits the use of the pan pot as a means to achieve image localisation within stereophonic recordings. As such, only an interaural intensity difference exists between left and right channels for a single source. We use gain scaling and phase cancellation techniques to expose frequency dependent nulls across the azimuth domain, from which source separation and resynthesis is carried out. The algorithm is demonstrated to be state of the art in the field of sound source separation but also to be a useful pre-process to other tasks such as music segmentation and surround sound upmixing.

Declaration:

I certify that this thesis which I now submit for examination for the award of PhD is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. This thesis was prepared according to the regulations for graduate study by research of the TU Dublin and has not been submitted in whole or in part for another award in any other third level institution. The work reported on in this thesis conforms to the principles and requirements of TU Dublin's guidelines for ethics in research. TU Dublin has permission to keep, lend or copy this thesis in whole or in part, on condition that any such use of the material of the thesis be duly acknowledged.

Signature _____ Date _____

ACKNOWLEDGMENTS

It has been a long road and many people have helped me greatly along the way. Firstly, I would like to thank my supervisor, Dr. David Dorran, for convincing me to revisit this thesis after a lengthy hiatus. His support, guidance and friendship has been critical to the completion of this document over the last year 12 months. I would also like to thank Dr. Eugene Coyle and Dr. Bob Lawlor who acted as my supervisors when I originally started my postgraduate research.

Many thanks to my good friend, Dr. Ted Burke, for proof reading this thesis and for many enlightening conversations and brainstorms over the years. Thanks also to Dr. Mikel Gainza for his great friendship and insight along the way on our journey through academia and startups together. I would also like to thank Dr. Derry Fitzgerald for many enjoyable conversations on sound source separation and music. Thanks also to Mr. Frank Duignan who helped me port the ADRess algorithm to Java many years ago. Thanks also to Dr. Gavin Kearney for his help on subjective testing.

I would also like to thank my friends and family for all the support along the way. Finally, I would like to thank my wife, Ali, for all the love and support throughout this whole adventure. This thesis is dedicated to our little boy, Alex.

TABLE OF FIGURES

ABOUT THE AUTHOR

My deep interest in audio signal processing is driven by my passion for music performance and production. Since my early teens I have played guitar, bass and drums in numerous bands and was lucky enough to tour Europe and the U.S. with my first band Bambi in the late 90s.

After my undergraduate education in Electronic Engineering, I undertook a taught Masters in Music and Media Technology in Trinity College to learn more about the technology and algorithms behind music production. This began my research journey, which eventually led me to be the Vice President of Research and Development with Fender Musical Instrument Corporation after they acquired my company Riffstation, which I formed with my research colleague Dr. Mikel Gainza in 2011.

Riffstation is often described as "Guitar Hero for Real". The application is built upon several audio signal processing algorithms, one of which is the Azimuth Discrimination and Resynthesis (ADRess) algorithm which I successfully patented while I was undertaking my initial PhD research in 2003. Since its release, Riffstation has been downloaded by millions of guitar players worldwide and has received strong recognition from the music community.

While a full-time PhD researcher, I established the DIT Audio Research Group in 2006 with two of my research colleagues. I became the manager of the group and operated as PI on our research projects. Under my management, the group won €2.1 million in funding for various projects from Enterprise Ireland, European Framework

FP6 and Science Foundation Ireland. The group grew to 15 researchers at its peak and published over 80 publications in various conferences and journals, and ultimately led to the formation of Riffstation in 2011.

While my commitments to the Audio Research Group and Riffstation were a barrier to submitting my PhD dissertation at an earlier stage, I was able to develop many other aspects of my research skill set. I acted as principal supervisor on two successfully completed MPhil projects and as secondary supervisor for one PhD. I have commercialized my research, having licensed the ADRess algorithm to Sony for use in the game SingStar which has sold over 13 million units. I also have 21 publications, 381 citations, 3 patents, and acted as reviewer for the AES, IEEE and was a member of the organising committee for the IET Signals and Systems Conference.

CHAPTER 1: INTRODUCTION

Sound Source Separation refers to the task of extracting individual sound sources from some number of mixtures of those sound sources. As an example, consider the task of listening in humans. We have two ears which means that our auditory cortex receives two sound mixtures, one from each ear. Through some complex neural processing, the brain is able to decompose these mixtures into perceptually separate auditory streams. A well known phenomenon known as the "Cocktail Party Effect" (Cherry, 1953) illustrates this process in action. In the presence of many speakers, humans exhibit the ability to tend to or focus on a single speaker despite the surrounding environmental noise. In the case of music audition we exhibit the ability to identify the pitch, timbre and temporal characteristics of individual sound sources within an ensemble music recording. This ability varies greatly from person to person and can be improved with practice but is present to some degree in most people. Even young children whilst singing along to a song on the radio are carrying out some form of sound source separation in order to discern which elements of the music correspond to a singing voice and which do not.

In engineering, the same problem exists. A signal is observed which is known to be a mixture of several other signals. The goal is to separate this observed signal into the individual signals of which it consists of. This is the goal of this research. In particular, this research is concerned with separating individual musical sound sources from ensemble music recordings for the purposes of audition, analysis, transcription, segmentation, remixing and upmixing.

Song = bass + guitar + drums + piano + voice

Stated simply: observing only the mixture(s) of these instruments - i.e. the song - the aim is to recover each individual sound source or instrument present in the song.

1.1 - APPLICATIONS OF SOUND SOURCE SEPARATION

There is quite literally a multitude of applications where sound source separation could be utilised, here are but a few that appear in the literature.

Music Education: A common problem for amateur musicians is identifying exactly which instrument is playing which note or notes in polyphonic music. A sound source separation facility would allow a user to take a standard musical recording such as a song on a compact disc, and extract an individual instrument part. Inversely, a single instrument may be muted. A tool such as this is a valuable asset in both the teaching and learning of music. For instance a music student would be able to extract an instrument of his/her choice in order to analyse and learn that musical part. Or conversely, the student could remove an instrument so that he or she would be able to play their part along with the remaining accompaniment.

Music Transcription: Transcription is the process of transforming some set of audio events into some form of notation. In the case of music, it typically involves creating a musical score from audio. Traditionally, this task was carried out by humans and was both expensive and laborious. Computerised music transcription tools have expedited the process but are generally limited to either monophonic transcription or a

special case of polyphonic transcription whereby the overall musical harmony (notes from all instruments grouped together) can be transcribed, but accuracy is still well below that of a human expert (Benetos et al. 2013). The inaccuracies are then corrected by a human using a suitable editing interface (Lunaverus, 2019). Sound source separation can aid this process by allowing a polyphonic mixture to be decomposed into several monophonic mixtures thus allowing established transcription techniques to be applied.

Music Composition: Computerised compositional tools are available at little cost to the user now. Integrated software and hardware packages have made it possible for a single desktop computer to contain almost all of the functionality of a commercial recording studio. At the time of publishing the major contributions of this work, sound source separation was not an established tool in music composition software. However, such features have become more common since 2004.

Audio Analysis: In many real-world scenarios, audio recordings can often be corrupted by unwanted noise from sound sources which are proximal to the source of interest. Forensic audio analysis is one such example. Source separation would facilitate the isolation of particular sounds of interest within corrupted recordings.

Remixing and Upmixing: Multichannel audio formats such as the Dolby and DTS 5.1 surround sound formats have become a standard in the film industry. More recently, multichannel spatial audio formats such as Ambisonics have been adopted for use in virtual and augmented reality. Upmixing is the process of generating several

reproduction channels out of only one or two mixtures. Using sound source separation, old films and music, for which the multitrack recordings are unavailable, could be remastered for today's common formats.

1.2 - ORGANISATION OF DISSERTATION

This dissertation is being submitted as part of the requirements for the award of PhD, under TU Dublin's academic regulations. Unlike a traditional dissertation, this is a PhD by prior publication, and as such is organised differently. Chapter 1 presents an overview of the document including the novel contributions and the applications of sound source separation. Chapter 2 is a review of the prior art at the time the novel contributions were made. Chapters 3, 4, 5, 6 and 7 present my novel contributions in the field. Each of these chapters comprises a previously published paper in its entirety with no edits, although the text has been reformatted for the purpose of maintaining document consistency. As such, each novel contribution chapter has its own internal structure containing an abstract, background, method and results section. Further, the introductions to each novel contribution chapter may overlap in background content. Chapter 8 presents conclusions and future work. All of the references are presented at the end of this document.

1.3 - NOVEL CONTRIBUTIONS

Within this dissertation, I present one major novel contribution in Chapter 3 and four secondary contributions within chapters 4, 5, 6 and 7 respectively. The key novel contribution is the Azimuth Discrimination and Resynthesis algorithm (ADRess) which is presented in Chapter 3. The algorithm was first published in 2004 and has

since been cited 177 times between its two published papers and one US patent. The patent has been cited by Sony, Samsung, Dolby and NEC. The algorithm was licensed to Sony in 2006 for use in SingStar on the Sony PlayStation 3, which went on to sell 13m copies. In 2012, the algorithm was licensed to Riffstation, a company I co-founded, which went on to be acquired by Fender Musical Instruments Corporation where it served millions of users globally from 2012 to 2018.

The second novel contribution is presented in Chapter 4 and explores two alternative methods of reconstructing the sources separated using the ADRess algorithm. The paper presented in this chapter, “Comparison of Signal Reconstruction Methods for the Azimuth Discrimination and Resynthesis Algorithm”, explores *Sinusoidal Modelling* and *Magnitude only Reconstruction* as alternatives to the original reconstruction method presented in Chapter 3.

The third novel contribution is presented in Chapter 5 and explores a novel use of the ADRess algorithm to achieve Music Structure Segmentation. In this chapter we show that an intermediate representation created by ADRess, the *azimugram*, can be further processed using Independent Subspace Analysis to segment musical audio into contextual sections such as verses and choruses.

The fourth contribution is presented in chapter 6 and explores a novel way of using the ADRess algorithm to upmix from stereo to a 5 channel surround presentation. Here, ADRess is configured to produce 5 fully reconstructable audio stems to serve as independent channels in a surround sound mix. Objective and subjective testing are

used to compare the stereo upmix generated surround mixes against true surround mixes of the same content.

The fifth and final contribution is presented in Chapter 7 in which a single channel drum source separation algorithm is presented. The algorithm was originally designed to overcome a shortcoming of the ADRess algorithm. Specifically that of the case where multiple sources are panned to the same azimuth, in which case ADRess cannot separate them. The drum separation algorithm was designed as a post process for ADRess but it was also shown to be a very useful preprocess for Prior Subspace Analysis-based drum transcription algorithms.

Chapters 3 - 7 of this dissertation are based on the following publications:

1. Barry, D., Lawlor, R., Coyle, E. (2004). Sound Source Separation: Azimuth Discrimination and Resynthesis. 7th International Conference on Digital Audio Effects, DAFX 04. Naples, Italy. October 5-8.
2. Barry, D., Lawlor, R., Coyle, E. (2005). Comparison of Signal Reconstruction Methods for the Azimuth Discrimination and Resynthesis Algorithm. 118th Audio Engineering Society Convention. Barcelona, Spain. May 28-31.
3. Barry, D., Gainza, M., Coyle, E. (2007). Music Structure Segmentation using the Azimugram in conjunction with Principal Component Analysis. 123rd Audio Engineering Society Convention. New York, USA. October 1.
4. Barry, D., Kearney, G. (2009). Localization Quality Assessment in Source Separation-based Upmixing Algorithms. 35th Audio Engineering Conference. Audio for Games. London. February 1.
5. Barry, D., Fitzgerald, D., Coyle, E., Lawlor, R. (2005). Drum Source Separation using Percussive Feature Detection and Spectral Modulation. IEE Irish Signals and Systems Conference. Dublin, Ireland. Sep 1-2.

The novel contributions, designs and implementations of the algorithms from the publications listed above are my work alone. However, I'd like to acknowledge Prof. Eugene Coyle and Dr. Bob Lawlor for acting as supervisors, and Gavin Kearney for helping conduct subjectives tests in his dedicated facility.

CHAPTER 2: LITERATURE REVIEW

Sound source separation as a field of study spans over many general topics in the wider fields of signal processing, machine learning, cognitive psychology and the physiology of hearing. In this literature review, I introduce the foundational concepts which led to the novel contributions presented later in this document.

2.1 - COMPUTATIONAL AUDITORY SCENE ANALYSIS

Auditory scene analysis (ASA) (Bregman, 1990) refers to the way in which the human auditory system is capable of decomposing concurrent sounds impinging on the ears, into a set of perceptually separate sound events despite the fact that the individual sounds may overlap considerably in both the time and frequency domains. Bregman, a psychologist by profession, conducted experiments using human listeners who were subjected to various audio tests, from which a set of conclusions were drawn about human organisation of sound. The purpose of the testing was to identify what mechanisms we as humans use in order to perform sound source separation. In other words, how do we 'know' which time and frequency components of a sound mixture belong to which of the individual sounds in the mixture? Two forms of such organisation were identified, simultaneous and sequential organisation.

- Simultaneous organisation deals with the separation and grouping of sounds occurring at the same time, this corresponds to grouping across the frequency domain.
- Sequential organisation is responsible for grouping similar components which occur at different times, the simplest example being that of a melody in a song.

A tenet of gestalt visual grouping psychology (Palmer 2003) known as *common fate* is the basis of Bregman's simultaneous organisation concept. Common fate refers to the notion that components behaving in a similar fashion are most likely related in some way. If several components are seen to have similar frequency or amplitude modulation characteristics, it could be inferred that they are related in some way. Figure 2.1 shows two synthetic sources playing simultaneously, one with vibrato (frequency modulation) and one without. No harmonics are overlapping in the example. It is quite easy to identify which harmonics belong to which source.

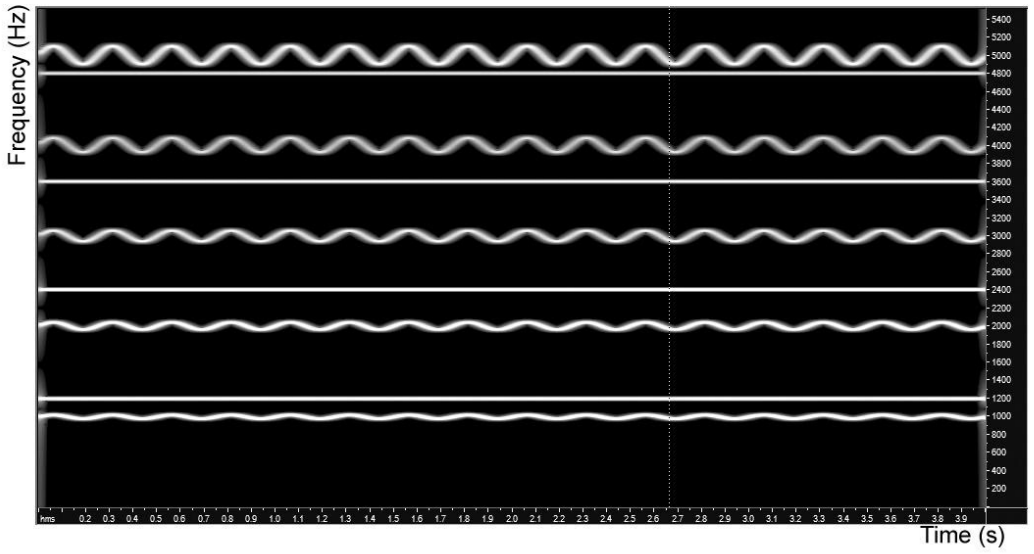

Figure 2.1 Spectrogram of two sources with fundamentals at 1000 Hz and 1200 Hz. Source one contains five harmonics and frequency modulation at a depth of 20Hz and rate of 4Hz. Source two contains four harmonics and no frequency modulation. The modulated source is clearly visible.

In a similar way it would be possible to identify frequency components with amplitude co-modulation. Common onset and offset of components are also an element of the common fate concept. If a new sound enters a mixture, it will contribute energy to the existing mixture at frequencies where it has energy. In a

similar fashion, a sound offsetting or leaving the mixture will result in decaying energy in regions of the spectrum where it had influence. The concept of harmonicity is another simultaneous grouping mechanism. A set of frequency components are said to be harmonically related if they fall into a pattern whereby each component is an integer multiple of some fundamental frequency. In human binaural listening, the spatial location of a sound will further reinforce grouping since all frequencies emanating from a single sound source will originate in the same location.

Sequential organisation deals with grouping sound as time evolves. Bregman refers to this as 'perceptual streaming'. Successive sounds with similar spectra are likely to form a perceptual stream. The spatial location of a sound is also a sequential grouping method. In the sequential case, however, it is the sounds emanating from the same location at different times which are perceptually grouped. Bregman also suggests that differences in intensity and phase may account for perceptual streaming.

This review of Bregman's work is by way of background to the area of Computational Auditory Scene Analysis (CASA) (Brown et al. 1994) and (Ellis, 1996) which, as the name suggests, uses much of the ASA research carried out by Bregman. CASA systems use these perceptually motivated heuristics as the basis for computerised sound source separation. In effect CASA attempts to model the way in which we as humans carry out the separation task. Furthermore these systems often incorporate prior knowledge of instrument characteristics and music composition rules to aid separation. In a similar way, humans become more familiar with the sound of certain

instruments due to repeated exposure. With this in mind we move to the most commonly occurring processes within a CASA system.

At the outset, it may seem like a simple task to computerise ASA because as humans, we all have an innate ability to separate sources. However, when it comes to specifying the problem to such an extent that it may be programmed as an algorithm on a computer, simple human concepts turn into significant engineering problems. As an example, timbre is a word used to describe the perceptual quality of a musical instruments' sound. 'It sounds bright, and shrill' is a reasonably generic description of how a human might describe the sound of a trumpet but descriptors like bright and shrill are not easily quantifiable. They are merely verbal descriptions of a percept and so a direct translation to computerisation is not simple. In (Ellis, 1992), Ellis states, "*Probably the hardest part of any complete source separator will be the simulation of the functions served by memory and experience in human listeners. It is not clear how well we would be able to organise and segregate composite sounds if we did not already have a good idea of the character of the individual sources based on previous examples*". Let us now consider some of the major building blocks of a CASA system.

2.2 - ONSET DETECTION

An onset can be defined as the point in time at which a new audio event enters the sound mixture. Onsets are often referred to as transients or attacks; both terms convey specific meaning but should not be used interchangeably even though the attack portion of a note may be transient in its nature. Duxbury makes this distinction by

saying that a note onset characterises the start of a new sound object whereas a percussive transient refers to a burst of noise (Duxbury, 2003). He suggests then that percussive transients become a subset of note onsets. This distinction is valid since the schemes to detect each type of onset vary significantly. A more detailed explanation suggests that a perceptual onset can be defined as '*the perceived beginning of a discrete event, determined by a noticeable increase of intensity*' or by '*a sudden change in pitch or timbre*' (Moelants et al. 1997). Onset detection is usually the first step in any transcription system (Klapuri, 1998). In (Klapuri, 1998), onset detecion is used to locate the presence of audio events first, then pitch discrimination and musical grouping heuristics are applied. A simple form of onset detection can be attained by first differentiating and rectifying a time domain signal followed by applying some form of envelope tracking technique. Convolution with a hanning window is usually the method applied or a simple low-pass operation. An onset candidate is then considered as anything which exceeds a given threshold. Peak picking is then used to locate the instant of the onset. Figure 2.2 illustrates the action of the onset detector. Equation 2.1 produces the envelope plot seen in figure 2.2 and equation 2.2 is a simple thresholded peak picker which produces the onset plot.

$$E(t) = H(t) * |x'(t)| \qquad (2.1)$$

where $|x'(t)|$ is the absolute value of the derivative of the time domain signal, $H(t)$ is a suitable hanning window and * denotes convolution. $E(t)$ is the resulting energy envelope.

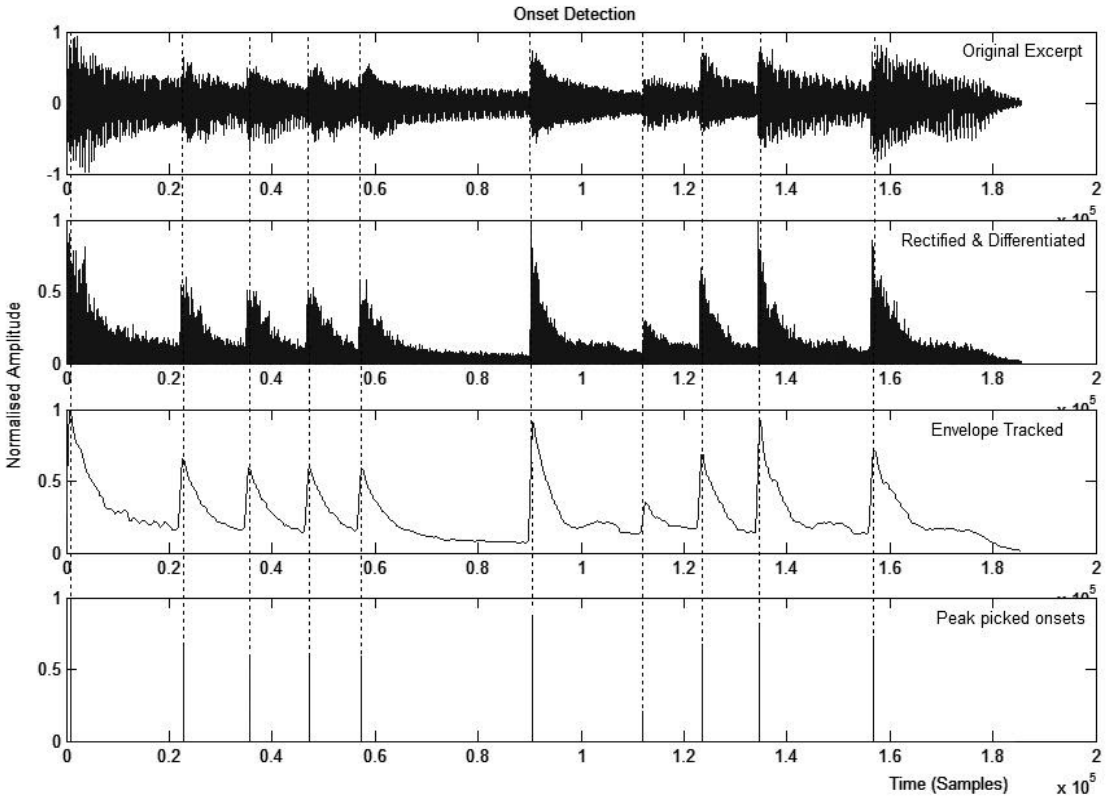

Figure 2.2 shows the detection of onsets within a short monophonic piano excerpt.

$$O(t) = E(t) \quad for: \quad E(t\text{-}1) < E(t) > E(t+1) \quad and \quad E(t)_t > Threshold$$

$$O(t) = 0 \; elsewhere \tag{2.2}$$

O_t is the resulting onset plot. This is a rudimentary form of onset detection used for illustration purposes. It would be reasonably well suited to the detection of transients but less effective for the detection of softer onsets such as those produced by bowed string and some wind instruments. As such, many variations of onset detection exist, some more successful than others (Bello et al. 2005). Frequency-domain onset detectors as opposed to time domain approaches such as that described above have both advantages and disadvantages. They allow more complex onset detection schemes which can be localised within certain frequency bands or even discontinuous

groups of bands. The main disadvantage of these systems is that the time resolution is significantly reduced due to the time windowing requirements of time-frequency transforms such as the short time Fourier transform (STFT) (Allen, 1977). In (Masri, 1996) such an approach is presented. This approach is based on the idea that during the attack portion of a note, i.e. the onset, an increase in high-frequency energy is usually observed. This is especially true of hard transient-like onsets but not necessarily the case for softer onsets such as those of certain wind instruments. The general technique involves applying a biased linear weighting function to the short-time magnitude spectrum of the audio. The biasing is in favour of high frequencies. It is achieved simply by multiplying the square of the magnitude of the k^{th} frequency bin by the bin number k. Equations 2.3, 2.4 and 2.5 describe this action.

$$E = \sum_{k=2}^{N/2+1} \left\{ | X(k) |^2 \right\} \tag{2.3}$$

where E is the computed energy of the fourier frame $X(k)$ of length N samples.

$$HFC = \sum_{k=2}^{N/2+1} \left\{ | X(k) |^2 \times k \right\} \tag{2.4}$$

where HFC means 'high frequency content'. Simply multiplying each bin magnitude by its bin index inherently gives more weight to the higher frequencies resulting in an energy measure, HFC, which relates more to the presence of high frequencies. Masri's approach takes the energy from two consecutive frames into account when deciding

whether or not an onset is present. An onset is then considered present if the following condition is met, equation 2.5.

$$\frac{HFC_r}{HFC_{r-1}} \times \frac{HFC_r}{E_r} > T_D \qquad (2.5)$$

where the subscript r represents the frame number and T_D is a threshold above which an onset is detected. Both (Scheirer, 1998) and (Klapuri, 1998) went on to develop multiband onset detection systems. Their schemes involve partitioning the audio into 6 sub-bands. In Scheirer's case, each sub-band is one octave wide. The amplitude envelope of each is extracted and smoothed. The first order difference function is then calculated, from which a local rise in energy in each band is gauged. Klapuri does much the same except he uses the relative difference function in order to more accurately locate the position of the onset. An onset candidate is then considered as anything exceeding a threshold. The actual onsets are then found as the candidates exhibiting the largest magnitude within a 50 ms sliding time window. Both of these approaches address the issue of soft onsets in that onsets are localised within bands and can be observed independently of energy present in other bands. Furthermore, it provides a starting point for pitch estimation in the case of soft onsetting instruments. Aside from these energy-based approaches there are also phase-based approaches such as that of Bello and Sandler (Bello, 2003). The phase vocoder is a well known technique for audio manipulation such as time scale modification (see section 2.2). Based on the short-time Fourier transform (STFT), the phase vocoder uses the phase information of two consecutive frames of audio separated by some distance referred to as the hop size to calculate what the phases for the next frame should be. Stated

simply, for a sinusoidal component of frequency f at time τ and with phase Φ, it will be possible to estimate what the phase at time $\tau + t$ will be. Of course, this is only the case if the sinusoid continues to be present into the next frame. An onset with energy at frequency f will be observed as a discontinuity in the sinusoidal track. Using this logic it would be possible to detect onsets based on the difference between the actual phases of a frame and the estimated phases of the same frame. If the two do not fall within a reasonable range of each other, then it could be ascertained that an onset occurred.

Although not a complete review, the above approaches represent the foundation of any onset detection scheme as it may be employed within a CASA system (Ellis, 1996). Assuming the onsets have been detected correctly, it would then be necessary to use some form of pitch detection in order to begin to separate out individual sources. This will be dealt with in greater detail in section 2.5 and a review of general pitch detection techniques can be found in (Ryynanen, 2004). In Chapter 7, novel work is presented to achieve drum source separation using a spectral-based onset detection method.

2.3 - BINAURAL PROCESSORS

Binaural processors, sometimes referred to as ‘cocktail party processors’, aim to perform sound source separation based principally on the localisation cue. The localisation cue is responsible for our ability to know ‘where’ a sound is coming from in the physical space around us; the cocktail party effect illustrates this. This cue is particularly powerful when considering the problem of sound source separation for

music. Most CASA systems are concerned with tracking the ever-changing pitches of instruments which come and go throughout the duration of a composition, thus making grouping a very difficult task. On the other hand, the position of a musician on stage rarely changes. Even during artificial playback a single source will usually remain in the same position throughout the length of a song. With this in mind, research has been carried out on the possibility of separating sound sources based primarily on their spatial location. The main localisation cues are interaural intensity difference (IID) which is predominant for frequencies above 1.5 kHz and interaural time difference (ITD) which is predominant at frequencies lower than 1.5 kHz. The ITD cue gives rise to what is known as the Haas effect, sometimes called the precedence effect (Haas, 1972). This psychoacoustic phenomenon refers to the fact that reflections of a sound impulse which occur within 30-40 ms of the direct sound will be perceptually fused as one event. This in turn gives rise to the law of 'the first arriving wavefront' which refers to the fact that a sound will generally be localised at the position corresponding to the origin of the direct sound. The angle of incidence of a sound, often called azimuth, is then derived from the time of arrival difference at each ear. This time difference is as a direct result of the fact that the path length of a single source will be different for each ear unless of course the source is directly in front or 0 degrees in the lateral plane. In a similar way, IIDs contribute to localisation due to the fact that a sound will be perceived as being louder in the ear which is closest to the source. This is true due to the inverse square law which would indicate that a longer path length will result in greater attenuation of the sound. This is further affected by head shadowing effects which will further attenuate frequencies with wavelengths less than the dimension of the human head. These phenomena would

suggest that the localisation cue could form the foundations of a robust sound source separation system. (Blauert, 1998) presents such a system. Most binaural models attempt to simulate the effects of the outer, middle and inner ear. The outer ear, called the pinna is particularly difficult to model due to the fact that they vary from person to person (Middlebrooks et al. 1991). In fact the left and right pinnae on a human head show slight differences. This too aids localisation, since the folds in the cartilage will accentuate and attenuate certain frequencies depending on the angle of incidence of sound. In general, the response of the pinna is usually modelled as a direction dependent linear filter (Blauert, 1998). The middle ear is usually just modelled as a bandpass filter in the range 20 Hz - 20 kHz. The inner ear and in particular the basilar membrane is modelled here as a bank of adjacent bandpass filters with critical bandwidths (Plomp, 1965). The neural excitation pattern caused by hair cell firing on the basilar membrane is simulated by rectifying and lowpass filtering (800 Hz cut off) the bandpass signals. The firing intensity in each critical band is then proportional to the obtained time functions. The binaural processor in this case is implemented as a cross-correlation between the left and right inputs of the system. This is done for each band. The maximum output of the cross-correlation function then corresponds to the time lag of either the left or right input. Also included in this model is a mechanism called ‘contralateral inhibition’ which attempts to model the Haas effect by suspending the system output for some milliseconds after a steep onset is encountered. In this way, reverberant reflections are suppressed and false directional information is omitted. The output from each processing band is considered to be what Blauert calls a binaural excitation pattern. The results of the cross correlation for each band can be converted to azimuth, after which the bands are weighted and

summed. The output of this model, a binaural excitation pattern, does indeed show the lateral displacement of sources but it offers no actual source separation as such. Furthermore, localisation in this model is based on the ITD cue which is valid for real-world listening, and in fact the inputs to this system are derived from dummy-head recordings in real environments, but in the case of music it is likely that the signals requiring separation have come from a recording studio and such signals rarely have discernable ITDs. (Roman, 2001) extended this research to produce a system capable of segregating speech from a noise mixture. In this instance, the inner ear was modelled using a 128-band gammatone filter. The bands are then weighted in accordance with the equal loudness curves (Fletcher, 1933). The output of each band is then processed in much the same way as (Blauert, 1998) using half wave rectification to simulate firing probabilities of the nerves. The azimuth locator was again based on the cross-correlation technique (Jeffress, 1948) except that in this model the function is limited to a range of ±1 ms since the maximum possible delay will correspond to the width of the human head. This model is extended by the formation of a binary mask based on the ITDs extrapolated from the cross correlation and the energy ratios for each signal. For frequency components below 1.5 kHz the binary mask is set to 1 when the ITD for a given frame and frequency channel exceeds a threshold and for frequencies above 1.5 kHz when the energy ratio exceeds a certain threshold. This binary mask is then applied to the channel in which the source of interest has greatest magnitude. The result is that the outputs of the gammatone filter bank are ‘switching’ on and off as the binary mask evolves through time.

This may be acceptable for the case of a speech noise mixture but in the case of music, there will typically be multiple sources present overlapping in both time and frequency. In this case the simple 'all or nothing' binary mask would not suffice in the case where the magnitude in a single frequency bin may be the sum of several instruments contributing energy at that frequency. A similar model is presented in (Bodden, 1996) and an overview of binaural models can be found in (Stern, 1985).

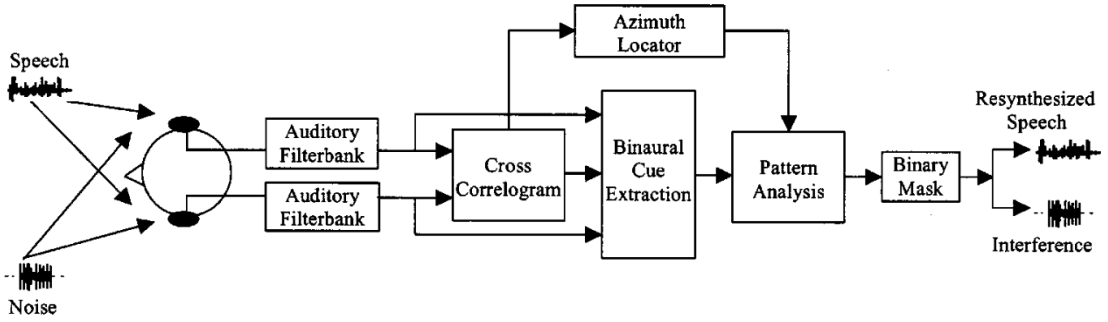

Figure 2.3 Schematic diagram of Roman's model. Binaural signals are obtained by convolving input signals with head-related impulse responses (HRIR). A model of the auditory periphery is employed. Azimuth localization for all the sources is based on a cross-correlation mechanism. ITD and IID are computed independently for different frequency channels. A pattern analysis block produces an estimation of an ideal binary mask, which enables the reconstruction of the target signal and the interfering sound.

Binaural separation models form the foundations of the novel contributions presented in Chapter 3.

2.4 - SINUSOIDAL MODELING AND THE PHASE VOCODER

Sinusoidal modelling is a well known technique for the analysis and synthesis of harmonic signals such as speech and music. First proposed by McAulay and Quatieri (McAulay, 1986), the technique describes how such signals can be represented as the sum of a set of quasi sinusoidal waveforms and a noise component each with time varying characteristics. The sinusoidal part of the signal is referred to as deterministic

whilst the noise part is considered as stochastic. Both the deterministic and stochastic elements of speech and music can be considered stationary over short periods of time. The assumption then is that the deterministic part of a sound can be resynthesised by extracting the instantaneous amplitudes, phases and frequencies from short time frames after which the values can be interpolated to form 'tracks' which can be resynthesised. The stochastic part is then found by subtracting the deterministic part from the original signal resulting in a noise residual. This is then modelled as white noise convolved with a time varying filter after which it is added back to the deterministic signal to yield a representation of the original sound. Prior to resynthesis, the parameters can be manipulated in order to achieve some desired effect such as time scale modification (Flanagan et al. 1966) or indeed sound source separation such as in (Virtanen, 2002). The model for sinusoidal modelling can be described by:

$$x(t) = \sum_{n=1}^{N} a_n(t)\, cos(2\pi f_n(t) + \Phi_n(t)) \qquad (2.6)$$

where $a_n(t)$, $f_n(t)$ and $\Phi_n(t)$ represent the amplitude frequency and phase of the n^{th} harmonic at time t. $r(t)$ is the stochastic or residual noise part of the signal and can be described as:

$$r(t) = \int_0^t h(t,\tau)u(\tau)d\tau \qquad (2.7)$$

where $u(\tau)$ is white noise and $h(t,\tau)$ is the response of a time varying filter to an impulse at time *t*. In other words the residual is modeled by convolving white noise with a time varying filter.

The process starts with a frequency analysis such as that of the STFT (Allen, 1977) which results in a short time phase and magnitude spectrum for each STFT frame. Strong sinusoidal components will be seen as peaks in the magnitude spectrum. A peak is considered as any bin with a magnitude greater than that of its two nearest neighbours. Peaks below a certain threshold are discarded and considered as belonging to the stochastic part of the signal. Once the peaks have been identified, parameter estimation occurs which involves estimating the frequency amplitude and phase of the peak. Due to the time-frequency resolution trade-off as a result of the STFT, the amplitudes and frequencies of the peaks are usually estimated by fitting a parabola to the 3 bins around a peak. The true maximum of this function is then taken to be the amplitude and its position is used to calculate the true frequency of the sinusoid. The phase can be derived directly from the phase spectrum of the frame. Once the parameters for each frame have been calculated, a peak-continuation algorithm attempts to link peaks from frame to frame resulting in a set of partial tracks. Generally, the algorithm tries to link each peak in a frame to a corresponding peak in the next frame. Linking occurs if the frequency of a peak in the next frame lies within a certain range of the frequency of a peak in the current frame. If a suitable match is not found, the track is said to have 'died' and its amplitude is set to zero in the next frame. Subsequently, there will be 'new' peaks in the next frame which have

not been matched to peaks in the previous frame, these tracks are said to have been 'born' and the amplitudes of those peaks in the previous frame are set to zero. For the resynthesis, linear interpolation is used for the amplitudes while cubic interpolation is used for the phase for each partial track after which all components are summed as in equation 2.8.

$$d(t) = \sum_{n=1}^{N} a_n(t)\, cos(2\pi f_n(t) + \Phi_n(t)) \qquad (2.8)$$

where $a_n(t)$, $f_n(t)$ and $\Phi_n(t)$ represent the amplitude frequency and phase of the n^{th} harmonic at time t. $d(t)$ represents the deterministic part of the signal. The stochastic signal, $r(t)$, is synthesised such that:

$$r(t) = x(t) - d(t) \approx \int_0^t h(t,\tau)u(\tau)d\tau \qquad (2.9)$$

where $x(t)$, in this case, is the original signal. The deterministic part of the signal, $d(t)$, is subtracted from $x(t)$ to give $r(t)$ which is then modeled by convolving white noise with a suitable filter. An alternative method for resynthesis involves creating complex arrays filled with the amplitude and phase parameters corresponding to each partial frequency after which an IFFT is used to generate short time frames corresponding to the original audio. The frames would need to be overlapped in accordance with the method used during the analysis stage. This method, although faster, is significantly less accurate since the parameter values are fixed for the duration of a frame thus resulting in a quantised time resolution. Figures 2.5 and 2.6 show the analysis and synthesis process of sinusoidal modelling (Serra, 1997).

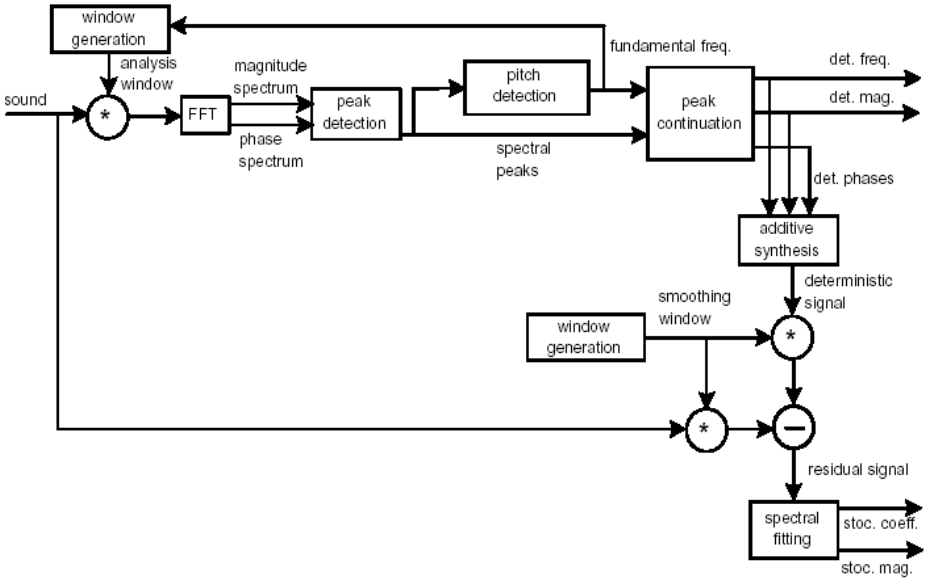

Figure 2.4 Schematic diagram of the analysis section of Serra's sinusoids + noise model. Image reproduced from (Serra, 1997)

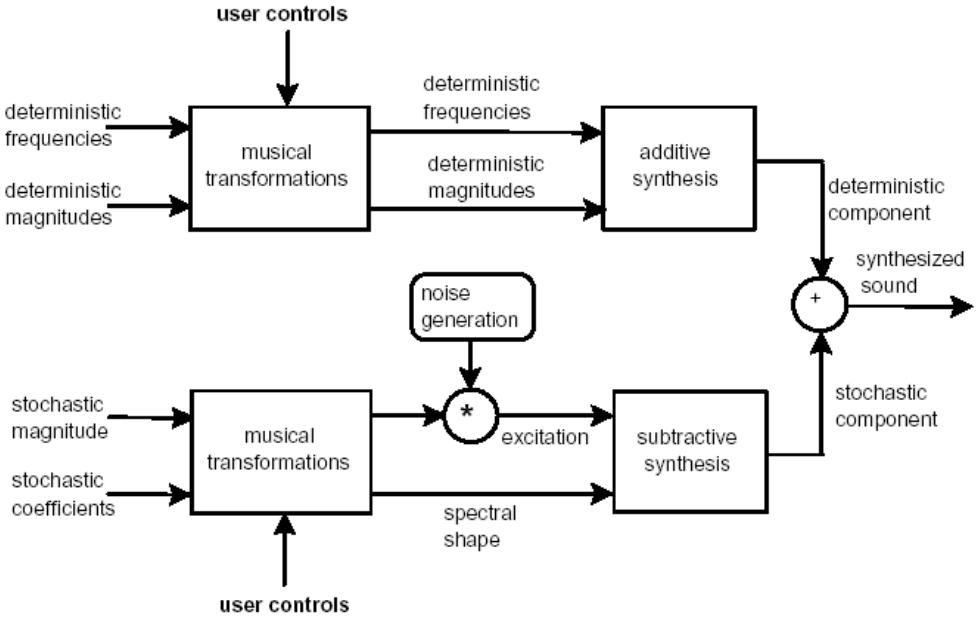

Figure 2.5 Schematic diagram of the synthesis section of Serra's sinusoids + noise model. Image reproduced from (Serra, 1997)

Referring to figure 2.4, the analysis process starts with a windowed fast fourier transform from which magnitudes and phases are estimated. Peak detection is then carried out followed by pitch detection. The fundamental frequency in turn informs the window generator at the input. Next a peak continuation algorithm generates sinusoidal tracks comprising of deterministic frequencies, magnitudes and phases. Additive synthesis is used to create the deterministic signal which is subtracted from the input signal to calculate the stochastic or residual signal. This signal is then modeled as spectrally filtered white noise. Figure 2.5 shows how the deterministic and stochastic coefficients from the analysis stage can be subjected to musical transformations before resynthesis.

In Chapter 4, I explore the use of sinusoidal modelling as an alternative method for reconstructing sources separated using the ADRess algorithm (Barry et al. 2005).

2.4.1 - The Phase Vocoder

The phase vocoder (Flanagan, 1966) in itself is not generally associated with sound source separation but elements of it have appeared in such applications as the phase-based onset detector in (Bello, 2003). Furthermore the phase vocoder provides an introduction to the STFT. The short-time Fourier transform operates on a time-domain signal and produces a time-frequency representation of that signal. If parameters are chosen correctly, an inverse Fourier transform may reproduce the original signal faithfully. In order to obtain the STFT of a signal, it is first broken up into short time frames of N points, usually in the order of 10-100 ms in length, where $t_s=N/Fs$ (time in seconds equals the number of sample points divided by the sample

frequency). A discrete Fourier transform (DFT) is then applied to each frame. This results in an *N*-point complex frequency array, the absolute value of which represents the magnitude spectrum of the current analysis frame. This is done successively for each frame. In order to obtain any information about the frequency content of a signal, the frame length must be greater than 1 and typically is greater than 256 to acquire any reasonable frequency resolution for a nominal sample frequency of 44.1 Khz. Commonly, the frame length is a power of 2 which allows for the computationally efficient Fast Fourier Transform (FFT) to be used instead of the DFT. The frequency resolution will rise as a function of frame length, but the time resolution decreases. This results in a tradeoff between time and frequency resolution, since good frequency resolution is required to distinguish close frequency components and good time resolution is required to encapsulate rapid changes in the time domain. One partial solution to this problem involves overlapping the analysis frames which corresponds to having an analysis step size, usually called the hop size, which is less than the length of the frame. A 50% overlap with a 4096-point frame relates to a hop size of 2048 points. This means that the last 2048 points of the first frame is the same as the first 2048 points of the second frame. On resynthesis this will produce amplitude modulation. The solution to this problem is to multiply the time-domain frame by a windowing function such that the overlapping portions of the frame always sum to 1. The Hanning window is such a window. Windowing serves a second purpose. Rectangular windows will usually have discontinuities at the start and the end of the window, this leads to the presence of high frequency components in the frequency transform which are not actually present in the signal. The Hanning window or any raised cosine window for that matter causes the signal to be faded in

gradually and faded out again, thus avoiding any discontinuities at the frame boundaries. The side effect of this is that the energy is slightly smeared in the frequency domain leading to attenuated wider main lobes and the presence of side lobes. However, the smearing is still less significant than using a rectangular window. The Hanning window is described by equation 2.10 and the DFT is given by equation 2.11

$$h(n) = 0.5\,(1 - cos(2\pi\tfrac{n}{N-1})) \qquad n=0,...,N-1 \qquad (2.10)$$

where N is the window length and n is an index into N.

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j\Omega kn} \qquad (2.11)$$

where $\Omega k = 2\pi k/N$ and Ωk is the angular frequency in the k^{th} frequency bin.

The resulting *X(k)* contains a complex frequency array. Only the first *N/2* points are required due to the fact that anything above that point corresponds to frequencies above the Nyquist (*Fs/2*). The magnitude spectrum can be obtained as $|X(k)|$ and the wrapped phases can be obtained from $\angle X(k)$. In the context of the phase vocoder, the STFT is given as:

$$X(t_a^u, \Omega_k) = \sum_{n=-\infty}^{\infty} h(n)x(t_a^u + n)e^{-j\Omega_k n} \qquad (2.12)$$

where x is the original signal, *h(n)* is the windowing function and $\Omega_k = \frac{2\pi k}{N}$ is the centre frequency of the k^{th} vocoder channel in radians per sample. $t_a^u = uRa$ where u

is the frame number and *Ra* is the analysis hop size. One use of the phase vocoder is to carry out time scale modification of audio (Barry et al. 2008). The phase vocoder does this by using a resynthesis hop size (*Rs*) different to that of the analysis. If the *Rs* >*Ra*, the audio will be time scale expanded and vice versa. Either the analysis or the synthesis hop size may be varied in order to achieve time scale modification. The time scaling factor is calculated as: α=*Rs/Ra* . In order for the new time scaled frames to overlap synchronously, the frame phases must be updated according to the phase propagation formula in equation 2.13.

$$\angle Y(t_s^u, \Omega_k) = \angle Y(t_s^{u-1}, \Omega_k) + R_s \hat{\omega}_k(t_a^u) \qquad (2.13)$$

where $\hat{\omega}_k(t_a^u)$ is the instantaneous frequency given by equation 2.14.

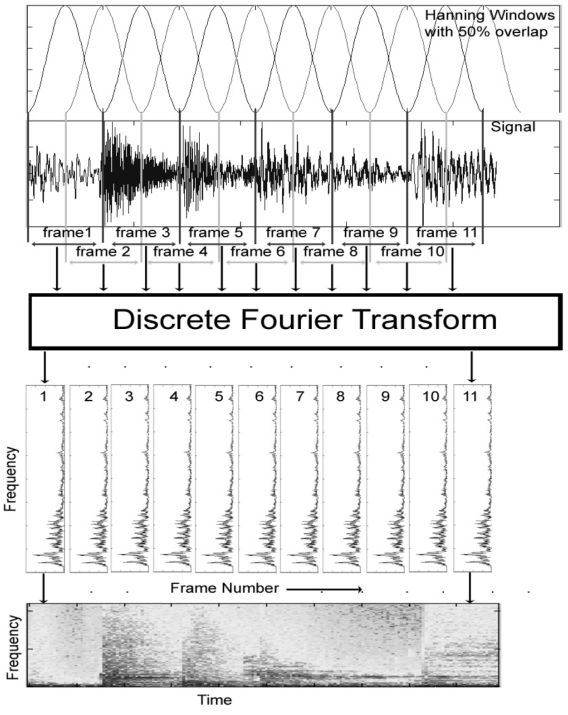

Figure 2.6 The STFT – short overlapped time frames are multiplied by a suitable hanning window after which a DFT is carried out on each resulting in a time-frequency representation of the audio.

$$\hat{\omega}_k(t_a^u) = \Omega_k + \frac{1}{Ra}\Delta_p\Phi_k^u \qquad (2.14)$$

where $\Delta_p\Phi_k^u$ is the principal argument of the heterodyned phase increment given by:

$$\Delta_p\Phi_k^u = \angle X(t_a^u, \Omega_k) - \angle X(t_a^{u-1}, \Omega_k) + R_a\Omega_k \qquad (2.15)$$

The new updated phases for each synthesis frame are given by equation 2.13 and the magnitudes are simply obtained by setting $\left|Y(t_s^u, \Omega_k)\right| = \left|X(t_a^u, \Omega_k)\right|$ where $t_s^u = R_s u$. Resynthesis is then carried out using an inverse Fourier transform on each frame using the new phase and magnitude values along with the synthesis hop size, $R_{s,}$ instead of the analysis hop size, $R_{a.}$

2.5 - STATISTICAL METHODS

Several statistical methods have gained popularity in the field of blind source separation beginning in the early 2000s (Hyvarinen, 2000), (Smaragdis et al. 2003), (Fitzgerald, 2004). Originally referred to as statistical signal processing or information theoretic approaches, these techniques are now more commonly classed as machine learning techniques and more specifically unsupervised learning approaches. The most common approaches used in blind source separation are discussed in the following sections.

2.5.1 - Independent Component Analysis

Independent component analysis (ICA) is a statistical method for discovering the latent variables which underlie some observable data which is a mixture of such latent variables (Hyvarinen, 2000). For example, imagine four different mixtures of four different people speaking. Theoretically, ICA should be able to recover each individual speaker given only the four mixtures. However, in the ICA model, each mixture is assumed to be linear and non-convolutive. This means that each individual source should be phase coherent across all mixtures and that each individual source is subjected to the same convolution conditions within any mixture. The mixing model can then be defined as:

$$x = As \qquad (2.16)$$

where $x = (x_1, ..., x_m)$ is a matrix of observed mixtures and $s = (s_1, ..., s_n)$, is the unknown matrix of independent components or sources. A is an $m \times n$ invertible matrix called the mixing matrix which is also initially unknown. The idea is to find an 'un-mixing' matrix W such that:

$$y = Wx = WAs \qquad (2.17)$$

where $y = (y_1, ..., y_n)$. This matrix y should contain the independent components of x assuming that the variables are non-gaussian and mutually independent. Variables are considered statistically independent if and only if the product of their marginal

densities is equal to the joint density of the same variables as in equation 2.18 (Hyvarinen, 2000)

$$P(y_i, y_j) - \left[P(y_i)P(y_j)\right] = 0 \qquad i \neq j \qquad (2.18)$$

ICA requires that at least as many observation mixtures as sources are present in order for ICA to successfully separate each source. In the case of consumer music media, there are generally only two-channel mixtures corresponding to the left and right channels of a stereo mix. This effectively means that ICA is limited to separating only mixtures containing at most two linearly mixed sources. Furthermore, the independent components are randomly ordered and usually scaled by some unknown factor. ICA is best suited to blind source separation problems where the observed data have been acquired using multi-sensor arrays unlike the case of musical sound source separation where typically there are only 2 observation mixtures. In (Barry et al. 2005 b), an attempt is made to overcome the limitation of needing at least as many observation mixtures as sources present. In that paper, standard ICA techniques were applied to contiguous magnitude frames of the short-time Fourier transform of the mixture. Provided that the amplitude envelopes of each source are sufficiently different, it can be seen that it is possible to recover the independent short-time power spectrum of each source. A simple scoring scheme based on auditory scene analysis cues is then used to overcome the source ordering problem ultimately allowing each of the independent spectra to be assigned to the correct source. A final stage of adaptive filtering is then applied which forces each of the spectra to become more independent. Each of the sources is then resynthesised using the standard inverse short-time Fourier

transform with an overlap add scheme. The algorithm was capable of source separation in very limited cases.

ICA was also applied to the task of music transcription with promising results in (Abdallah et al. 2003). In this case, the limitation of needing as many sensors as sources was overcome by considering the spectrogram to be the sum of individual note spectra which were assumed to be sparse (mainly zero entries in the matrix). As we will see in the following sections, these techniques often perform better at the task of transcription than source separation. It should be noted that the two overlap considerably in terms of the approaches taken. In general, these statistical techniques tend to be good at modeling note spectra and therefore good at transcription, but less capable of attributing the detected notes to the instrument of origin which would be required for instrument separation.

2.5.2 - Principal Component Analysis

Principal component analysis (PCA) is a dimensionality reduction technique which is sometimes referred to as eigenvalue decomposition or singular value decomposition (SVD). In the case of a matrix, the model assumes that the information contained within that matrix can be represented by lower dimensional subspaces, the sum of which approximates the original matrix. Each subspace is the result of the outer product of a latent basis function, *W,* a vector of dimension $m \times 1$ and a time activation function, *H,* a vector of dimension $n \times 1$, where $m \times n$ is the dimension of the original data **X,** a matrix. Formally stated, it is assumed that the matrix **X** can be decomposed into a sum of outer products as in equation 2.19.

$$\mathbf{X} = \sum_{j=1}^{J} \mathbf{V}_j = \sum_{j=1}^{J} W_j H_j^T \qquad (2.19)$$

where T denotes the transpose of the matrix. In matrix notation, $\mathbf{X}$ is represented as the sum of J subspaces $\mathbf{V}_j$, each one corresponding to a particular latent "feature" of the original data.

These basis functions are obtained by carrying out singular value decomposition, more commonly known as PCA, on the matrix. This essentially transforms a high-dimensional set of correlated variables into some number of lower dimensional sets of uncorrelated variables which are known as the principal components. The principal components are ranked in order of variance, so the first principal component contains the maximum amount of total variance present in the data and each subsequent principal component represents the maximum remaining variance in the data.

In the context of source separation for musical applications, our audio signal is first represented as a matrix. A time frequency representation such as the spectrogram is typically used in the literature. In (Fitzgerald et al. 2002), the spectrogram of a signal which contains a mixture of drums is represented by $\mathbf{X}$ using the notation of equation 2.19. They postulate that $\mathbf{X}$ can theoretically be represented as the sum of J independent spectrograms, $\mathbf{V}_j$, where each one contains a single drum (kick, snare, hat etc.). Discovering the independent spectrograms, $\mathbf{V}_j$, directly is a difficult task but

applying PCA to the mixture spectrogram reduces the dimensionality of the data in some logical way. Given that the drums are pitch stationary, their individual spectra will be broadly similar throughout the duration of the drum hit and further, drum hits on the same drum should be similar in each case. As a result, it is generally the case that applying PCA to a drum mixture results in the discovery of a latent principal component and time activation function for each drum. Referring to equation 2.19, the principal components are represented by W_j and H_j. By getting the outer product of a single basis function and its time activation function, an approximate spectrogram for each drum can be constructed. Figure 2.7 shows the time activation functions and frequency basis functions obtained from a piece of music using Independent Subspace Analysis (ISA).

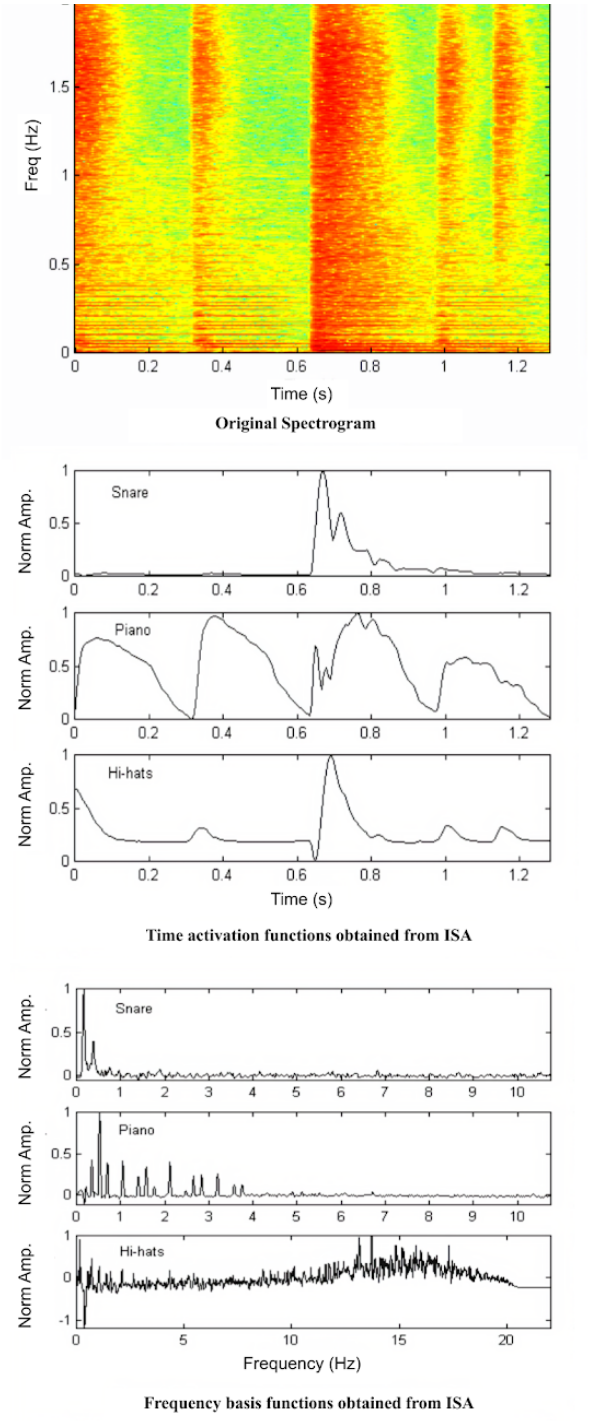

Figure 2.7 ***Top*** *- A spectrogram of a short song passage .* ***Middle*** *- The time activation functions obtained from ISA.* ***Bottom*** *- The frequency basis function obtained from ISA. (Fitzgerald 2004)*

2.5.3 - Independent Subspace Analysis

PCA on its own does not in general return a set of statistically independent basis functions as its purpose is to generate uncorrelated basis functions. As a result of both time and frequency overlap, the time activations for each drum may have a considerable amount of activity from other drums present. One way to address this is to perform ICA on the time activation functions. ICA optimises for independence, and therefore forces the time activations to be as independent as possible. Performing PCA followed by ICA is known as independent subspace analysis (ISA) and was first introduced by (Casey et al. 2000).

Another point of note within (Fitzgerald et al. 2002) is that the ISA method is performed on a sub-band basis. This helps with cases where two drums with minimal frequency overlap are hit at the same time. Take the case of a hi hat for example. Common beats will contain a hi hat strike at every down beat. This means there is a high probability that they will overlap with the kick and the snare on a regular basis.

One issue with using PCA or ISA is that of choosing how many principal components to use to represent the data (Fitzgerald et al. 2002). In the drum separation application, the number of components, J , is set to the expected number of recurring drums within the song. For example, if one expects to separate a kick, snare and hi hat, and those sources contribute to the most variance in the spectrogram, then a good place to start would be to set $J = 3$.

However, it is rarely the case that setting J equal to the number of expected sources will suffice (Fitzgerald et al. 2002). In general, some greater number of principal

components must be recovered to faithfully capture the audio characteristics of the underlying sources. To address this Fitzgerald extended the work in (Fitzgerald et al. 2003) where Locally Linear Embedding (LLE) is used instead of PCA within an ISA framework. There it is shown that LLE is able to characterise sources with fewer numbers of components than are required using PCA. This, according to Fitzgerald, is because LLE makes use of local geometry to embed high dimensional data in a lower dimensional space.

In practice, these techniques work well when separating individual drums within a mixture of drums but are not robust enough to reliably separate drums from polyphonic mixtures with the same accuracy. Furthermore, because of the pitch stationary limitation, the technique doesn't work well for polyphonic mixtures of pitched sources. As a result of these limitations, these techniques have been more successful at the task of transcription than at the task of separation. To that end, the work was extended further in (Fitzgerald et al. 2005) where a technique known as *prior subspace analysis* (PSA) was used to achieve pitched instrument transcription. The method can work with polyphonic instruments such as guitar and piano but in the case where there is more than one instrument playing, the algorithm is not able to attribute notes to specific instruments, instead giving the overall harmonic transcription. The transcription algorithm in (Fitzgerald et al. 2002) was further improved by using a novel drum source separation step as preprocess in (Barry et al. 2005) which is presented as a novel contribution in Chapter 7.

2.5.4 - Non-negative Matrix Factorisation

Non-negative Matrix Factorisation (NMF) is a numerical technique popularised by (Lee et al. 1999) but based on the work presented in (Paatero et al. 1994, 1997). It is used to decompose a matrix into subspaces based on the premise that the matrix is composed of the sum of underlying low rank matrices often called *parts* or *topics* depending on the application. Thus it is often referred to as a linear parts-based decomposition. Although similar to PCA in terms of the goal, it differs considerably in terms of its non-negativity constraint and computational method. This means that the model only allows for additive combinations, not subtractive combinations. In (Lee et al. 1999), NMF was applied to images of faces. There, they show that NMF performs a parts-based decomposition of the image such that the parts correspond to features of face such as eyes, ears, mouth etc. It has been shown to work considerably better than PCA for image decomposition (Lee et al. 1999). Figure 2.8 shows a comparison of face decomposition between PCA and NMF.

$$\mathbf{V} \approx \mathbf{W} \times \mathbf{H} \qquad (2.20)$$

In the case of sound source separation, a spectrogram is used as the input matrix. For the spectrogram **V,** of dimension $m \times n$, where each element of $\mathbf{V} \geq 0$, NMF decomposes it into two matrices **W** and **H** of dimension $m \times j$ and $j \times n$ respectively, where each element of $\mathbf{W} \geq 0$ and $\mathbf{H} \geq 0$ and where J is the desired rank of the factorisation. The NMF model is summarised in equation 2.20 above.

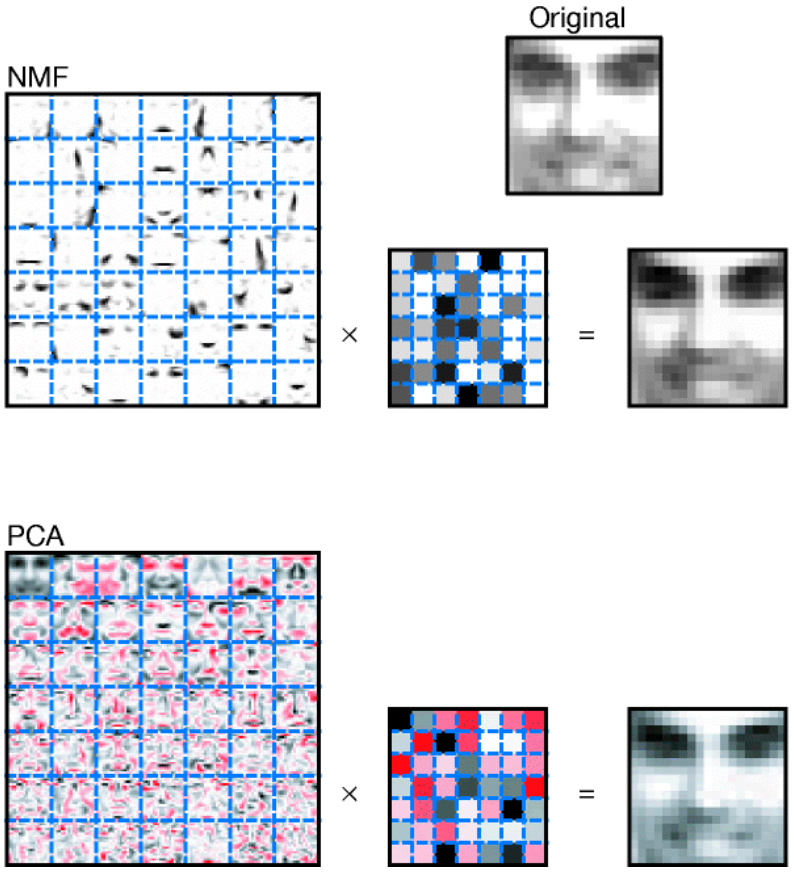

Figure 2.8. Adapted from (Lee et al. 1999) NMF learns a parts-based representation of a face but PCA learns a holistic representation. Looking at the NMF example, it can be seen that each learned feature closely resembles individual parts of a face such as eyes, nose and mouth. PCA on the other hand has learned how to approximate abstract variants of whole faces which when combined linearly approximate the target face.

Referring to Figure 2.9, the simplified spectrogram **V**, contains *n* time frames each containing *m* frequency bins. The decomposition then gives us **W**, which is a set of *J* frequency "parts" (similar to basis functions in PCA) where each part, represented in each column of **W**, models certain repeating characteristics of the music such as a specific note pitch, a specific pitch stationary instrument or a spectral feature of some kind. It should be noted that *J* refers to the rank which the user wishes to achieve. So if you expect to recover all instances of 72 unique pitched notes from the spectrogram you can expect to set *J* to at least 72.

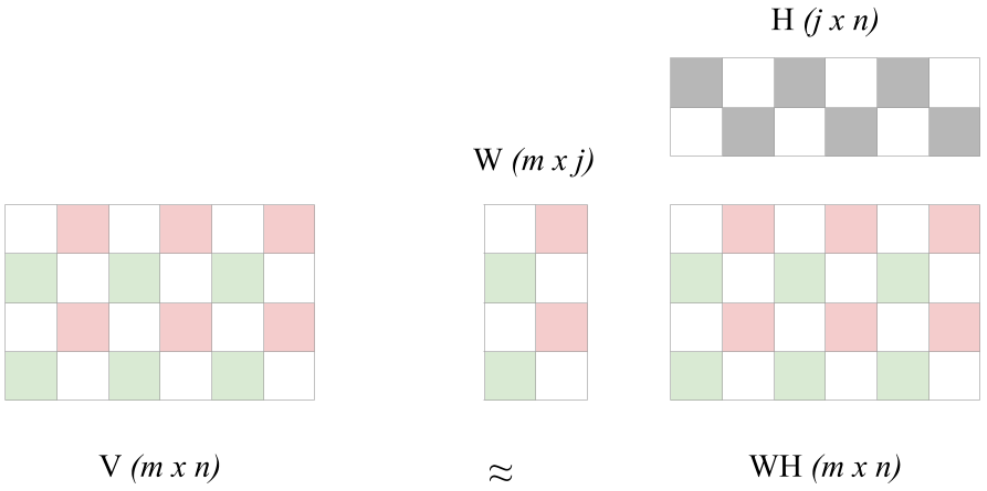

Figure 2.9. NMF decomposes a simplified spectrogram V into a set of spectral parts, W, and a set of time activation sequences for those parts

H is a set of J time activation sequences corresponding to each of the J spectral objects represented in **W**. Similar to PCA, the product of **W** and **H** approximates the original matrix. And similarly the outer product of $\mathbf{W}_{(m,j)}$ and $\mathbf{H}_{(j,n)}$ approximates the spectrogram of the J^{th} spectral object.

Obtaining the values for **W** and **H** involves a number of iterative steps:

1. Initialise **W** and **H** with random positive entries.
2. At each iteration, a suitable cost function is used to measure the distance or divergence between **V** and **WH**
3. A suitable update equation is used to update the values of **W** or **H** in each iteration such that the cost function is iteratively minimising. This amounts to non-negative linear regression.
4. Steps 2 and 3 are iterated until the cost function reaches a local minima.

The most common cost functions identified in (Lee et al. 2001) are the square Euclidean distance shown in equation 2.21 and the Kullback-Leibler divergence (KLD) shown in equation 2.22.

$$\|A - B\|^2 = \sum_{ij}(A_{ij} - B_{ij})^2 \qquad (2.21)$$

$$D(A\|B) = \sum_{ij}(A_{ij}\, log\frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij}) \qquad (2.22)$$

where **A** and **B** are **V** and **WH** respectively and *i* and *j* represent the column and row indices.

NMF can be stated as an optimisation problem with respect to equation 2.21 as follows:

Minimize $||\mathbf{V} - \mathbf{WH}||^2$ with respect to **W** and **H**, subject to the constraints that all elements of $\mathbf{W} \geq 0$ and all elements of $\mathbf{H} \geq 0$. Note that the function is convex only in **W** or **H** independently and not in both variables together and so only a local minimum will be found. This means that NMF may arrive at different solutions for the same problem on repeated factorisations depending on the random values set at matrix initialisation. Gradient descent provides a solution to arrive at local minima but may require several hundred iterations to converge. In (Lee et al. 2001) a multiplicative update method was proposed, shown in equation 2.23.

$$H \leftarrow H \frac{(W^T V)}{(W^T W H)} \qquad W \leftarrow W \frac{(V H^T)}{(W H H^T)} \tag{2.23}$$

where *T* denotes the transpose of the matrix

As mentioned above, a spectral object could be a note of a certain pitch or a pitch-stationary instrument such as a drum, but is unlikely to be a source in its own right. The reason for this is that all pitched instruments can produce pitched notes within their frequency range and many instruments overlap in range as shown in Figure 2.10. The same notes produced by different instruments will generally have the same harmonic relationships, i.e. integer multiples of the fundamental frequency, but the relative amplitude of those harmonics will be different for each instrument. Further, the dynamic variation of those harmonic amplitudes over time will differ between instruments. However, despite this, the parts-based decomposition nature of NMF in its raw form is far more predisposed to discovering notes than instruments or sources. For this reason it has been successfully applied to the task of polyphonic transcription than that of sound source separation (Smaragdis et al. 2003)(Fitzgerald et al. 2005 b).

Polyphonic transcription using NMF was first proposed in (Smaragdis et al. 2003). In Figure 2.11 below, the vertical plot on the left depicts the individual note spectra **W** and the bottom plot depicts the activations for each note spectrum in **H**. In this example *J = 4*.

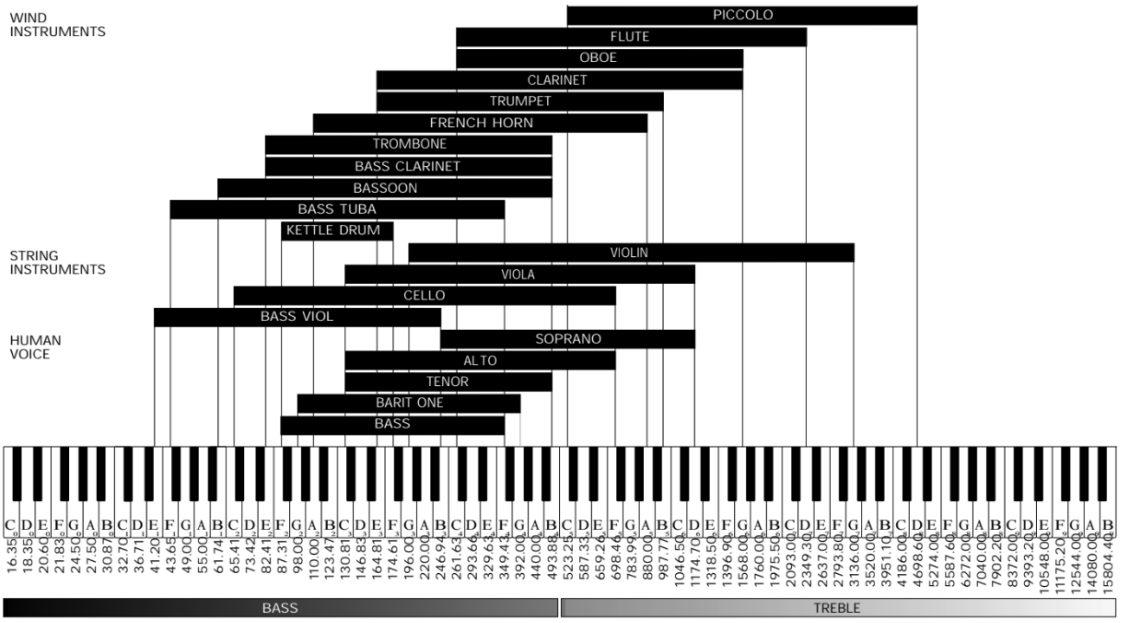

Figure 2.10. Pitch range of various instruments. Reproduced from James Husted, Symetrix http://educypedia.karadimov.info/library/PIANO.pdf

In (Fitzgerald et al. 2005 b), an NMF system to achieve instrument separation as opposed to note separation is presented. The approach overcomes the problem of attributing notes to the correct source by assuming that all notes belonging to a single source can be represented as a single frequency basis function which is translated to achieve different pitches.

This frequency basis function aims to capture the spectral characteristics of the source so that the same note played on different instruments can be attributed to the correct instrument. In order to be able to translate the basis function, a constant Q transform (CQT) is used instead of a spectrogram. Using a CQT allows for the logarithmic nature of the harmonic spectra to be translated as an integer shift in the matrix. Thus allowing for a single basis be used for all notes for a single instrument. The technique

overcomes the note grouping limitation of previous work but fails to produce results that merit further investigation (Fitzgerald et al. 2005 b). Figure 2.12 shows the separated spectrograms of a piano and flute using this technique.

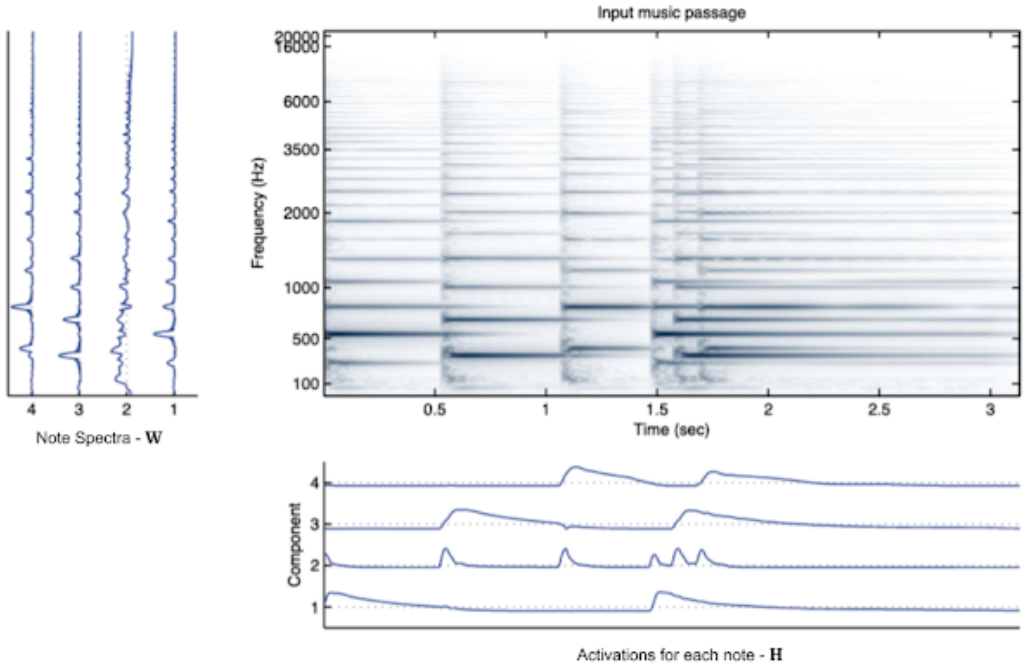

Figure 2.11: NMF decomposition of a polyphonic spectrogram (Smaragdis 2013)

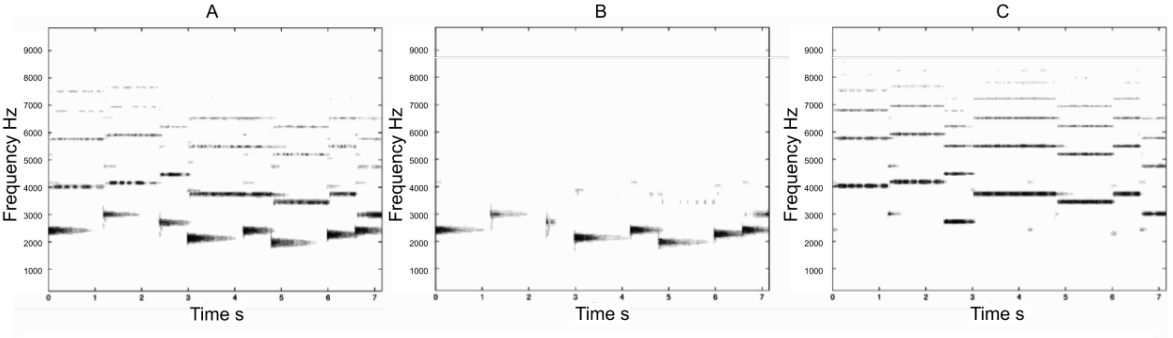

Figure 2.12 NMF-based separation of 2 source mixture. A: Mixture Spectrogram of Flute and Piano. B: Spectrogram of separated piano. C: Spectrogram separated flute. Reproduced from (Fitzgerald et al. 2005 b)

This work was extended further in (Fitzgerald et al. 2005 c) where the above technique was used on 2 channel mixtures instead of single channel mixtures. It is

based on the observation that the majority of commercial music is mixed to two channels, whereby a single source may exist in both channels with a different intensity in each. The process, known as *panning*, simply positions a source between the two channels by distributing the source to both channels using an intensity ratio. The work then utilises the fact that this intensity ratio can be used to group note spectra extracted using NMF. In other words, all notes with the same intensity ratio across both channels are assumed to derive from the same instrument. This improves upon the previous work but Fitzgerald concludes that objectionable artefacts still exist in the resynthesis as a result of the approximate nature of mapping the CQT log-frequency spectrograms back to linear-frequency spectrograms to allow for resynthesis. As we will see in the following sections, the concept of using 2-channel mixtures proves particularly useful when it comes to grouping separated frequency components by source.

2.6 - DEGENERATE UNMIXING ESTIMATION TECHNIQUE – DUET

The DUET algorithm (Jourjine et al. 2000) was designed for degenerate source separation of an arbitrary number of W-disjoint orthogonal (W-DO) sources using only two mixtures of those sources. Two sources are said to be W-disjoint orthogonal if the time-frequency representations of each source do not overlap significantly. It has been shown that this is approximately true in the case of speech (Rickard, 2002). This being the case, source separation can be achieved by creating a time-frequency binary mask for each source and applying it to the spectrogram of either mixture. The algorithm operates by estimating the amplitude ratio of each time-frequency point between the two mixtures and the time delay of each time-frequency point between

the two mixtures. The resultant delays and amplitude ratios are then used to create a two dimensional smoothed histogram where it can be seen that the values cluster into localised peaks as in Figure 2.13. These peaks are representative of source activity by virtue of the fact that for a single source remaining at the same location in space, we would expect to see the majority of its time-frequency points share the same amplitude ratios and delay coefficients. The mixing model for DUET can be defined as:

$$x_k(t) = \sum_{j=1}^{J} a_{kj} s_j(t - \delta_{kj}), \quad k = 1,2 \qquad (2.24)$$

where $x_k(t)$ is the k^{th} receiver mixture, a_{kj} and δ_{kj} are the attenuation coefficients and time delays related to the path from the j^{th} source to the k^{th} receiver for J sources. Moving to the time-frequency domain via the STFT we get:

$$X_k(m,n) = \sum_{j=1}^{J} (a_{kj} S_j(m,n) + \Delta\Phi_{kj}(m,n)), \quad k = 1,2 \qquad (2.25)$$

where $\Delta\Phi_{kj}(m,n)$ is a frequency dependent phase shift and where m and n are the time frame and frequency bin indices respectively.

For simplicity sake it is considered that the receiver closest to the j^{th} source is used as a 'reference' and so its amplitude coefficient can be set to 1 and delay coefficient set to 0. This is simply because we only need the inter-receiver time delay as opposed to the individual path length delays and similarly for the amplitudes. As a result we can represent $X_1(m,n)$ and $X_2(m,n)$ as:

$$X_1(m,n) = \sum_{j=1}^{J} S_j(m,n) \tag{2.26}$$

$$X_2(m,n) = \sum_{j=1}^{J} (a_j S_j(m,n) + \Delta\Phi_j(m,n)) \tag{2.27}$$

The inter-receiver attenuation coefficients for each bin are found using equation 2.28.

$$a_j(m,n) = \left| \frac{X_2(m,n)}{X_1(m,n)} \right| \tag{2.28}$$

The inter inter-receiver delay coefficients are then found using equation 2.29.

$$\delta_j(m,n) = -\frac{1}{n} \angle \left(\frac{X_2(m,n)}{X_1(m,n)} \right) \tag{2.29}$$

where $\angle$ implies the taking the phase angle in radians of the complex number resulting from the last term of equation 2.29.

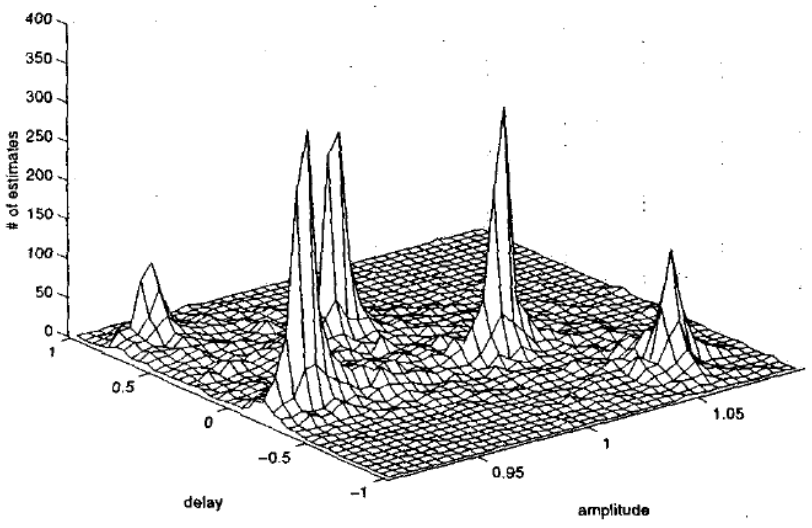

Figure 2.13 *2D Histogram showing 5 distinct peaks along normalised axes with different delay (δ_j) and attenuation (a_j) coefficients indicating the presence of 5 sources. Reproduced from (Jourjine et al. 2000)*

The result is that we now have both delay and attenuation estimates for each time-frequency point of the mixture. If each source is perfectly W-DO and the mixtures have been obtained under anechoic conditions, all frequency components belonging to one of the sources will have exactly the same attenuation and delay coefficients. This is rarely the case in the real world and what happens instead is that the parameter estimates obtained above are close to the ideal amplitude and delay coefficients. In order to identify the mixing parameters for each source, a 2-D smoothed and weighed histogram of delay against attenuation summed over time is created. Large peaks are seen in areas where several time-frequency points posses the same delay and attenuation coefficients. This is an indication that a single source relates to the mixing parameters which have caused the peak. Figure 2.13 shows such a histogram illustrating the presence of 5 sources. The coordinates of a peak in the histogram correspond to delay and attenuation coefficients which are present in many frequency channels.

Returning to the STFT, the source corresponding to a peak is extracted by creating a binary mask which sets all frequency bins with delay and attenuation coefficients, proximal to that of the peak, to '1' and all others to '0'. The mask for the j^{th} source can be defined as:

$$M_j(m,n) = \begin{cases} 1 & if \quad \left| A\Delta_j - \alpha\delta(m,n) \right| \leq R \\ 0 & otherwise \end{cases} \qquad (2.30)$$

where $A\Delta_j$ are the ideal attenuation and delay coefficients for the j^{th} source derived from a peak in the histogram. And where $\alpha\delta(m,n)$ are the attenuation and delay estimates for frequency bin *n* at time *m*; and R is a user defined range. Typical values for R would be 0.1 to 0.3 in the context of the normalised delay axis depicted in figure 2.13. If the estimates fall outside this range, those frequency components are set to 0. The source is then extracted by multiplying the binary mask by the original time-frequency representation as in equation 2.31:

$$S_j(m,n) = M_j(m,n)X_k(m,n) \qquad (2.31)$$

Experiments have shown that DUET is capable of very good results in separating an arbitrary number of speakers from only two mixtures obtained in real-world environments (Jourjine et al. 2000). It is the case that reverberant environments will deteriorate results significantly. There are limitations to the DUET algorithm: firstly the mixtures must be obtained from a pair of microphones which are no more than a few centimetres apart. In fact the maximum distance between the microphones is dependent on the frequency content of the signal being captured. The distance should be no greater than the wavelength of the highest occurring frequency within the signal. The reason for this can be attributed to the delay estimation technique which uses the phase difference between each mixture. The phase returned in equation 2.29 is a wrapped phase which is always in the range $+\pi$ to $-\pi$ radians, and so if any frequency component were to go through a full phase rotation before being received

at the lagging microphone, it would cause a delay estimate error as a result of phase wrapping. The distance between the mics will be limited by the following condition;

$$d = \delta_{j\max} c \qquad (2.32)$$

where

$$\delta_{j\max} = \frac{2\pi}{\omega_s} \qquad (2.33)$$

where c is the speed of sound in air and ω_s is the sampling frequency in radians per second.

The condition that sources must be W-DO for separation makes DUET unsuitable for the separation of musical signals since the basis of western tonal music is harmony. Harmony embodies the notion that when certain notes' spectra overlap, they produce richer combined spectra which can be pleasing to the ear. This is referred to as *tonal consonance*. It can be shown that the most consonant musical intervals correspond to the largest amount of frequency overlap (Howard, 2001). Attempts have been made such as that in (Master, 2003) to modify DUET for use with music but no results were presented.

2.7 - SOURCE SEPARATION IN LINEAR STEREO RECORDINGS

In (Avendano, 2003) a method for source identification and manipulation is described. The approach is based on the standard studio or artificial recording model which is largely linear. In a recording studio, each sound source is usually recorded

individually and summed across two channels with continuously varying intensity ratios. The model can be described by:

$$x_k(t) = \sum_{j=1}^{J} a_{kj} s_j(t), \quad k = 1,2 \tag{2.34}$$

where $x_k(t)$ is either the left or right mixture channel in the time domain, a_{ij} represents the attenuation factor for the j^{th} source in the k^{th} channel for each of J sources s_j .

Converting to the frequency domain using the STFT we get:

$$X_k(m,n) = \sum_{j=1}^{J} a_{kj} S_j(m,n), \quad k = 1,2 \tag{2.35}$$

where m and n are time frame and frequency bin indices respectively. A similarity measure of the input signals is used to identify time-frequency points occupied by each source based on the *panning coefficient* applied during mix down. The panning coefficient is similar to the amplitude coefficient in DUET; it is a ratio of energy between the left and right mixtures. Individual sources are identified and manipulated by clustering time-frequency components with similar panning coefficients. The similarity measure used is as follows:

$$\varphi(m,n) = 2\frac{\left| X_1(m,n)\, X_2^*(m,n) \right|}{\left| X_1(m,n)^2 \right| + \left| X_2(m,n) \right|^2} \tag{2.36}$$

where X_1 and X_2 are complex time-frequency domain representations of the left and right mixture channels respectively and where m and n are the time frame and frequency bin indices respectively. This similarity measure $\varphi(m,n)$ gives values proportional to the panning coefficients of each source provided that they do not overlap significantly in the time-frequency transform domain. This effectively means each source needs to have a different panning coefficient. The problem with this function is that it returns all positive values. This leads to "lateral ambiguity", meaning that the lateral direction of the source is unknown, i.e. a source panned 60° left will give an identical similarity measure to the one panned 60° right. To overcome this ambiguity, Avendano uses a partial similarity measure and a difference function defined respectively as,

$$\varphi_y(m,n) = 2\frac{\left|X_y(m,n)\,X_z^*(m,n)\right|}{\left|X_y(m,n)\right|^2}, \quad y \neq z \tag{2.37}$$

and

$$\Delta(m,n) = \varphi_1(m,n) - \varphi_2(m,n) \tag{2.38}$$

Now, positive values of $\Delta(m,n)$ correspond to sources panned towards the left and negative values correspond to sources panned towards the right. Values of zero correspond to non-overlapping time-frequency regions of sources panned to the centre. Avendano then defines an ambiguity resolving function as,

$$\widehat{\Delta}(m,n) = \begin{cases} 1 \;\; if \;\; \Delta(m,n) > 0 \\ 0 \;\; if \;\; \Delta(m,n) = 0 \\ -1 \;\; if \;\; \Delta(m,n) < 0 \end{cases} \tag{2.39}$$

The panning index map is then obtained by applying the above ambiguity resolving function to the similarity function as follows,

$$\Psi(m,n) = \left[1 - \varphi_1(m,n)\right] \widehat{\Delta}(m,n) \qquad (2.40)$$

Now time-frequency bins with similar panning indices can be clustered. In Avendano's case he uses a Gaussian window to scale bin magnitudes which are proximal to a specific panning index. A soft mask is constructed and applied to the short-time magnitude spectrum in order to separate a particular source. As with DUET, the bin magnitudes used for resynthesis are taken directly from the analysis STFT, thus the model assumes that all energy at a specific frequency corresponds to only one source. Furthermore the model is symmetrical in that sources which are panned by the same amounts in opposite directions will interfere with each other significantly. This effectively means that the separation quality deteriorates as the source moves away from the centre. Also of interest is that this approach uses absolutely none of the grouping heuristics of ASA, instead it takes advantage of the stereo audio format and the way in which stereo mixes are created.

The general approach of clustering frequency bins according to their IIDs, followed by binary masking of the clustered components within the STFT, has become a popular approach in recent years. In the 2007 Source Separation Evaluation Campaign (Vincent et al. 2007), all ten algorithms entered in the *instantaneous mixture* category used some variation IID or IPD clustering and STFT masking. In

(Bofill et al. 2001) the number of sources and the mixing coefficients are estimated from clustered peaks in an IID representation. STFT bins are selected as described in (Xiao et al. 2005). Source estimation is then achieved by minimising the l_1 norm of the real and imaginary parts of the source STFTs. In (Gowreesunker et al. 2007) peak picking and mixing coefficient estimation is carried out on a thresholded IID histogram. An MDCT is then used to resynthesise the source signals. The approach in (Kleffner et al. 2007) uses peak picking on a thresholded IID histogram in a similar fashion to (Mohan et al. 2003) and the STFT bins are selected similar to that of (Arberet et al. 2006). The source estimation is then carried out using minimum variance beamforming (Lockwood et al. 2004). In (Mitianoudis at al. 2007), it is assumed that the number of sources is already known. Here, ITD clustering is used in conjunction with an MDCT for source estimation. In (Vincent, 2007 b), manual peak picking was carried out on a weighted IID histogram similarly to (Arberet et al. 2006) followed by an l_0 norm minimisation of the source STFTs. The approach in (Xiao et al. 2005 b) also uses IID clustering but is designed to extract only 2 sources per time frame. In (Mandel et al. 2007), both IID and IPD clustering are used assuming the number of sources is known. Distance weighted masking is then applied to the STFTs to reproduce the sources. In chapter 3.10, ADRess is objectively compared with all of the algorithms which took part in the 2007 Source Separation Evaluation Campaign.

2.8 - REVIEW CONCLUSIONS

In general, the techniques outlined above have not been able to provide a complete solution to the problem of sound source separation for musical instruments. However, each has merits and forms a partial solution to the problem. The author notes that

CASA approaches are not wholly concerned with solving the sound source separation problem in the signal processing sense; instead the motivation is towards building systems which mimic the way in which the human auditory system works. The human analogy is often used for the sound source separation problem, whereby our ability to focus on particular sounds of interest is considered to be sound source separation. In actual fact, humans do not possess the ability to do sound source separation in any real sense. As an example, if many speakers are speaking in a room simultaneously, we cannot selectively hear only one speaker and suppress all interference from other speakers, we can however 'listen' to one speaker which involves focusing our attention. Hearing is a passive subconscious activity whereas listening is a conscious activity involving attention, memory and context. It would seem that the broadest goal of CASA is that of artificially intelligent machine listening. This said, many aspects of CASA research prove very useful for the problem at hand.

Of particular interest in CASA research is the binaural processor technique in which concurrent sounds are separated purely on the basis of their location in physical space. This model is immediately applicable to the general sound source separation problem. DUET (Jourjine et al. 2000) is effectively a realisation of the binaural model without the physiological modelling of the outer, middle and inner ear. It also seems to be one of the most effective techniques to date for sound source separation using only two sensors. The most significant limitation of DUET is the condition of W-disjoint orthogonality which states that the sources must not have significant overlap in the time or frequency domain. This makes DUET suboptimal for musical source separation since musical sources will by nature have significant amounts of overlap.

Statistical methods such as ICA are applicable only in cases where multiple mixtures can be observed. It is the case that the number of mixtures must equal the number of sources present in order for ICA to be successful.

The primary research carried out in this thesis concerns sound source separation of musical sources and further applications of the same. The goal is to successfully separate an arbitrary number of source signals from at most two observation mixtures corresponding to that of current musical media. The model presented in the next section finds its foundations in binaural processor techniques and draws on elements of ASA and DUET. The key difference is that the novel research presented in the coming chapters specifically optimises for the linear stereo recording mixing model.

CHAPTER 3: SOUND SOURCE SEPARATION: AZIMUTH DISCRIMINATION AND RESYNTHESIS

This chapter presents the principal novel contribution of this dissertation, the Azimuth Discrimination and Resynthesis algorithm (ADRess). It was originally published at the Digital Audio Effects Conference in 2004 (DAFX 04) and is presented here in its entirety in accordance with TU Dublin regulations (Barry et al. 2004). Sections 3.1 to 3.8 constitute the original publication, section 3.9 is additional work around real-time implementation and section 3.10 shows comparative test results against 10 other algorithms. The paper included co-authors Eugene Coyle and Bob Lawlor who acted as my PhD supervisors at the time. A second paper with some real-time additions published in the 117th Audio Engineering Society Convention proceedings later in 2004 (Barry et al. 2004 b) which is not reproduced here but the algorithm has since been cited 177 times between its two published papers and one US patent (Barry et al. 2011). The patent has been cited by Sony, Samsung, Dolby and NEC. The algorithm was licensed to Sony in 2006 for use in SingStar on the Sony PlayStation 3 which went on to sell 13m copies. In 2012, the algorithm was licensed to Riffstation, a company I co-founded, which went on to be acquired by guitar manufacturer Fender and served millions of users globally from 2012 to 2018. In 2019, the patent was licensed to VRX Audio which plans to use the algorithm as part of a spatial audio engine.

3.1 - ABSTRACT

In this paper we present a novel sound source separation algorithm which requires no prior knowledge, no learning, assisted or otherwise, and performs the task of separation based purely on azimuth discrimination within the stereo field. The algorithm exploits the use of the pan pot as a means to achieve image localisation within stereophonic recordings. As such, only an interaural intensity difference exists between left and right channels for a single source. We use gain scaling and phase cancellation techniques to expose frequency dependent nulls across the azimuth domain, from which source separation and resynthesis is carried out. We present results obtained from real recordings, and show that for musical recordings, the algorithm improves upon the output quality of current source separation schemes.

3.2 - INTRODUCTION

Our research is concerned with extracting sound sources from stereo music recordings for the purposes of audition and analysis. This is termed sound source separation and has been the topic of extensive research in recent years. In general, the task is to extract individual sound sources from some number of source mixtures. Currently, the most prevalent approaches to this problem fall into one of two categories, Independent Component Analysis, (ICA) (Hyvarinen et al. 2000),(Casey et al. 2000) and Computational Auditory Scene Analysis, (CASA) (Rosenthal et al. 1998). ICA is a statistical source separation method which operates under the assumption that the

latent sources have the property of mutual statistical independence and are non-gaussian. In addition to this, ICA assumes that there are at least as many observation mixtures as there are independent sources. Since we are concerned with musical recordings, we will have at most only 2 observation mixtures, the left and right channels. This makes pure ICA unsuitable for the problem where more than two sources exist. One solution to the degenerate case where sources outnumber mixtures is the DUET algorithm (Jourjine et al. 2000), (Rickard et al. 2001). Unfortunately this approach has restrictions which make it unsuitable for use with music. CASA methods on the other hand, attempt to decompose a sound mixture into auditory events which are then grouped according to perceptually motivated heuristics (Bregman, 1990), such as common onset and offset of harmonically related components, or frequency and amplitude co-modulation of components. We present a novel approach which we term Azimuth Discrimination and Resynthesis, (ADRess). The approach we describe is a fast and efficient way to perform sound source separation on the majority of stereophonic recordings.

3.3 - BACKGROUND

Since the advent of multichannel recording systems in the early 1960's, most musical recordings are made in such a fashion whereby *N* sources are recorded individually, then electrically summed and distributed across 2 channels using a mixing console. Image localisation, referring to the apparent position of a particular instrument in the stereo field, is achieved by using a panoramic potentiometer.

This device allows a single sound source to be divided into two channels with continuously variable intensity ratios (Eargle, 1969). By virtue of this, a single source may be virtually positioned at any point between the speakers. So localisation is achieved by creating an interaural intensity difference, (IID). This is a well known phenomenon (Rayleigh, 1907). The pan pot was devised to simulate IID's by attenuating the source signal fed to one reproduction channel, causing it to be localised more in the opposite channel. This means that for any single source in such a recording, the phase of a source is coherent between left and right, and only its intensity differs. It is precisely this that allows us to perform our separation. A similar mixing model is assumed in (Avendano et al. 2002) and (Avendano et al. 2003). It must be noted then, that our method is only applicable to recordings such as described above. Binaural, Mid-Side, or Stereo Pair recordings will not respond as well to this method although we have had some success in these cases also.

3.4 - METHOD

Gain-scaling is applied to one channel so that a source's intensity becomes equal in both left and right channels. A simple subtraction of the channels will cause that source to cancel out due to phase cancellation. The cancelled source is then recovered by creating a "frequency-azimuth" plane, which is analyzed for local minima along the azimuth axis. These local minima represent points at which some gain scalar caused phase cancellation.

It is observed that at some point where an instrument cancels, only the frequencies which it contained will show a local minima. The magnitude and phase of these minima are then estimated and an IFFT in conjunction with an overlap add scheme is used to resynthesise the cancelled instrument.

3.5 - AZIMUTH DISCRIMINATION

The mixing process we have described can be expressed as:

$$L(t) = \sum_{j=1}^{J} Pl_j S_j(t) \qquad (3.1a)$$

$$R(t) = \sum_{j=1}^{J} Pr_j S_j(t) \qquad (3.1b)$$

where S_j are the J independent sources, Pl_j and Pr_j are the left and right panning coefficients for the j^{th} source, and L and R are the resultant left and right channel mixtures. Our algorithm takes $L(t)$ and $R(t)$ as it's inputs and attempts to recover S_j, the sources. We can see from equation 3.1a and 3.1b that the intensity ratio of the j^{th} source, $g(j)$, between the left and right channels can be expressed as,

$$g_{(j)} = \frac{Pl_j}{Pr_j} \qquad (3.2)$$

This implies that $Pl_j = g(j).Pr_j$. So, multiplying the right channel, R, by $g(j)$ will make the intensity of the j^{th} source equal in left and right. And since L and R are simply the superposition of the scaled sources, then $L - g_{(j)}.R$ will cause the j^{th} source to cancel out. In practice we use, $L - g_{(j)}.R$, if the j^{th} source is predominant in the right channel and, $R - g_{(j)}.L$, if the j^{th} source is predominant in the left channel. This serves two purposes, firstly it gives us a range for $g(j)$ such that: $0 \leq g(j) \leq 1$. Secondly, it ensures that we are always scaling one channel down in order to match the intensity of a particular source, thus avoiding distortion caused by large scaling factors. So far we have only described how it is possible to cancel a source assuming the mixing model we have presented. In order to utilise this data, we move to the frequency domain. We divide the stereo mixture into short time frames and carry out an FFT on each,

$$Lf(k) = \sum_{n=0}^{N-1} L(n) W_N^{kn} \qquad (3.3a)$$

$$Rf(k) = \sum_{n=0}^{N-1} R(n) W_N^{kn} \qquad (3.3b)$$

where $W_N = e^{-j2\pi/N}$ and Lf and Rf are short time frequency domain representations of the left and right channels respectively. In practice we use a 4096 point FFT with a Hanning window and an overlap of 1024 points at a sampling frequency of 44.1 KHz. We create a frequency-azimuth plane for left and right channels individually as in

figure 3.1. The azimuth resolution, β, refers to how many equally spaced gain scaling values of g we will use to construct the frequency-azimuth plane. We relate g and β as follows,

$$g_{(i)} = (i).\frac{1}{\beta} \qquad (3.4)$$

for all i where, $0 \leq i \leq \beta$, and where i and β are integer values.

Large values of β will lead to more accurate azimuth discrimination but will increase the computational load. Assuming an N point FFT, our frequency-azimuth plane will be an $N \times \beta$ array for each channel. The right and left frequency-azimuth plane are then constructed using,

$$Az_{R(k,\,i)} = | Lf_{(k)} - g_{(i)}.Rf_{(k)} | \qquad (3.5a)$$

$$Az_{L(k,\,i)} = | Rf_{(k)} - g_{(i)}.Lf_{(k)} | \qquad (3.5b)$$

for all i and k where, $0 \leq i \leq \beta$, and $1 \leq k \leq N$.

It must be stated that we are using the term “azimuth” loosely. We are not dealing with angles of incidence. The azimuth we speak of is purely a function of the intensity ratio, created by the pan pot.

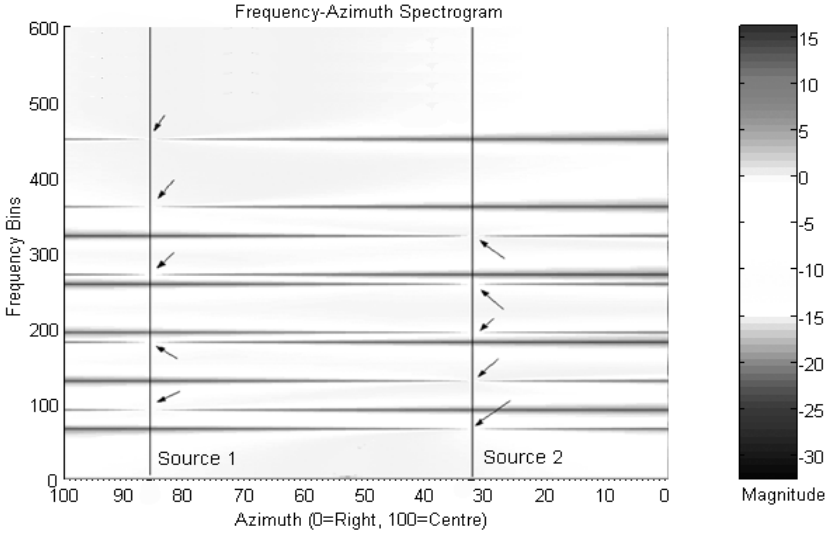

Figure 3.1: The Frequency-Azimuth Spectrogram for the right channel. We used 2 synthetic sources each comprising of 5 non-overlapping partials. The arrows indicate frequency dependent nulls caused by phase cancelation.

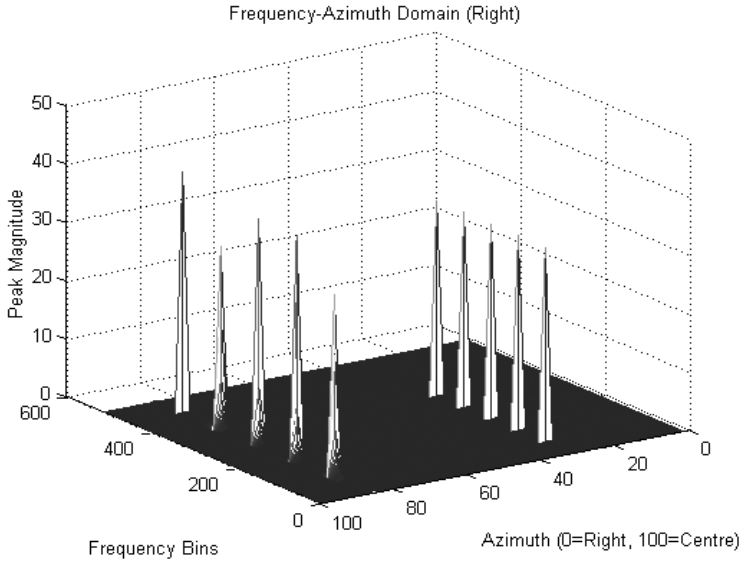

Figure 3.2: The Frequency-Azimuth Plane for the right channel. The magnitude of the frequency dependent nulls are estimated. The harmonic structure of each source is now clearly visible as is their spatial distribution. In order to estimate the magnitude of these nulls we redefine equation 3.5a and 3.5b:

In order to illustrate how this process reveals frequency dependent nulls, we generated two test signals, each with 5 unique partials. A stereo mix was created such that both sources were panned to the right, but each with a different intensity ratio. Using this test signal, the frequency-azimuth spectrogram in figure 3.1 was created using equation 3.5a, with, $\beta = 100$ and $N = 1024$. It can clearly be seen that frequencies have separated themselves out along the azimuth plane in figure 3.1 and 3.2.

$$AzR(k,i) = \begin{cases} AzR(k)_{\max} - AzR(k)_{\min} & \text{if } AzR(k,i) = AzR(k)_{\min} \\ 0, \text{ otherwise} \end{cases} \quad (3.6a)$$

$$AzL(k,i) = \begin{cases} AzL(k)_{\max} - AzL(k)_{\min} & \text{if } AzL(k,i) = AzL(k)_{\min} \\ 0, \text{ otherwise} \end{cases} \quad (3.6b)$$

Effectively, we are turning nulls into peaks as can be seen in figure 3.2. However, the test signal described, represents the ideal case where there is no harmonic overlap between 2 sources. This is almost never the case when it comes to tonal music. Harmony is one of the fundamentals of music creation, and as such instruments will more often than not be playing harmonically related notes simultaneously which implies that there will be significant harmonic overlap with real musical signals. The result of this is that frequencies will not group themselves as neatly across the azimuth plane as in figure 3.2. Instead, we observe *frequency-azimuth smearing*, whereby the frequency components from a single source will cluster loosely around a point in the azimuth plane as opposed to being perfectly positioned at precisely one point. This occurs when two or more sources contain energy in a single frequency bin.

The apparent frequency dependent null drifts away from a source position and may be at a minimum at a position where there is no source at all. For instance, if two sources in different positions, contained equal energy at a particular frequency, the apparent null will appear mid way between the two sources. It is the case that only sources predominant in the same channel will affect each other. A source in the left channel will not have an effect on a source in the right channel.

To overcome this problem, we define a user specified parameter, the *azimuth subspace width*, *H*, such that $1 \leq H \leq \beta$. This allows peaks within a given neighbourhood to be recovered. These azimuth subspaces can be overlapped if two sources are active in a single frequency bin. Peaks that drift away from their source positions can now be re-included for resynthesis. A wide azimuth subspace will result in worse rejection of nearby sources. On the other hand a narrow azimuth subspace will lead to poor resynthesis and missing frequency components. This parameter is varied by the user depending on the proximity of neighbouring sources. Figure 3.3 shows the same two test signals as before only each includes one extra partial of the same frequency. It can clearly be seen that the common frequency component is now apparent between the two sources. In order to recover it, the azimuth subspace boundary of the source must extend beyond it. This is shown for source one. At this point we introduce the "discrimination index", *d*. Where, $0 \leq d \leq \beta$. This index, *d*, along with the azimuth subspace width, *H*, will define what portion of the frequency-azimuth plane is extracted for resynthesis.

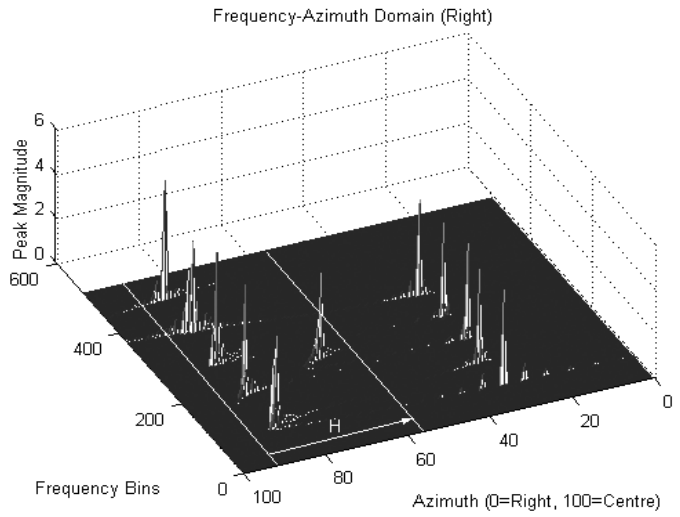

Figure 3.3: The Frequency-Azimuth Plane. The common partial is apparent between the 2 sources. The azimuth subspace width for source 1, H, is set to include the common partial.

3.6 - RESYNTHESIS

In order to resynthesise only one source, we set the discrimination index, *d*, to the apparent position of the source. In figure 3.3, there are 2 sources, one at approximately 85 points along the azimuth axis, and the other at 33. The azimuth subspace width, *H*, is then set such that the best perceived resynthesis quality is achieved. In practice, we centre the azimuth subspace over the discrimination index such that the subspace spans from *d-H/2* to *d+H/2*. The peaks for resynthesis are then extracted using,

$$YR(k) = \sum_{i=d-H/2}^{i=d+H/2} AzR(k,i) \quad 1 \le k \le N \tag{3.7a}$$

$$YL(k) = \sum_{i=d-H/2}^{i=d+H/2} AzL(k,i) \quad 1 \le k \le N \tag{3.7b}$$

The resultant *YR* and *YL* are *1 x N* arrays containing only the bin magnitudes pertaining to a particular azimuth subspace as defined by *d* and *H*. At this point it should be noted that, if two sources have the same intensity ratio, i.e. they share the same pan position, both will be present in the extracted subspace. This is particularly true of the "centre" position. It is common practice in audio mixing to place a number of instruments here, usually voice and very often bass guitar and elements of the drum kit too. In this instance, band limiting can be used to further isolate the source of interest.

The bin phases from the original FFT are used to resynthesise the extracted source, equation 3.8a and 3.8b. Once we have bin phases and magnitudes we can convert from polar to complex form using equation 3.9. The azimuth subspace is then resynthesised using the IFFT, equation 3.10.

$$\Phi_{R(k)} = \angle(Rf_{(k)}) \qquad (3.8a)$$

$$\Phi_{L(k)} = \angle(Lf_{(k)}) \qquad (3.8b)$$

Returning to the complex form using equation 3.9.

$$X_{(k)} = \left\{ \begin{matrix} \mathrm{Re}\, X_{(k)} = Y_{(k)}.\cos\Phi_{(k)} \\ \mathrm{Im}\, X_{(k)} = Y_{(k)}.\sin\Phi_{(k)} \end{matrix} \right\} \qquad (3.9)$$

We resynthesise our short time signal using the IFFT, equation 3.10

$$X_{(n)} = \frac{1}{N}\sum_{k=1}^{N} X_{(k)} W_n^{-kn} \qquad (3.10)$$

where $W_n = e^{-j2\pi/N}$

The resynthesised time frames are then recombined using a standard overlap and add scheme. Due to the fact that the magnitude spectrum for each frame and source is an estimate, the resynthesis is not perfect. The windowing function is not preserved and therefore the frames at the output do not tail off to zero as you might expect. As a result, some audible distortion may be present at the frame boundaries in the form of 'clicking'. This distortion arises from the fact that a small discontinuity will be present at the frame boundaries arising from the imperfect preservation of the window function. This was resolved by multiplying the output frames by a suitable windowing function which results in smoother frame transitions. The side effect of this solution is that an overlap of 75% must be used to avoid amplitude modulation in the output signal. It is intended that this algorithm will run in real time and that the control parameters d and H be set subjectively until the required separation is achieved. In effect, the user sweeps through the stereo space until the desired source is encountered. In much the same way as a pan pot places a source at some position between left and right, the ADRess algorithm will extract a source from some position between left and right.

3.7 - TESTING AND RESULTS

We have applied the ADRess algorithm to a number of commercial recordings. The degree of separation achieved depends on the amount of sources, the source proximity and the source level. If sources are proximate, it is likely that multiple sources may get extracted. If there is a large number of sources, partials may go missing. If the source level is too low, the resynthesis may have a bad signal to noise ratio. In general though, some degree of separation is possible. We generated a synthetic stereo signal, using 5 general midi instruments; bass, piano, drums, vibraphone and French horn. They were panned to 5 unique positions as in Figure 3.4.

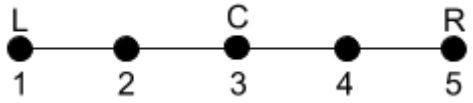

Figure 3.4: 5 sources panned to different positions. 1=bass, 2=vibraphone, 3=drums, 4=piano, 5=horn.

The piece of music in figure 3.5 was generated in a midi editor using these 5 instruments. The polyphony varies throughout the 2 bar segment with up to 9 notes sounding at once. In some cases 2 instruments are playing the same note at once. This poses no problem for ADRess since it depends only on a positional cue for separation. A stereo .wav file (figure 3.6) was then created using the score, instruments and panning parameters from above. This file was then processed by ADRess, with the relevant parameters set. The azimuth resolution, β, was set to 10 points for each side giving a total of 20 discrete pan locations between left and right. Higher values of β will give greater azimuth resolution but it is largely unnecessary since source

components spread out across the azimuth plane and will need an increased subspace width to recover the entire source. The azimuth subspace width, *H*, was set to 2 in all cases, corresponding to 20% of the entire azimuth subspace width of the channel being processed. The discrimination index, *d*, was set for each source position. A high quality of separation was achieved for all sources.

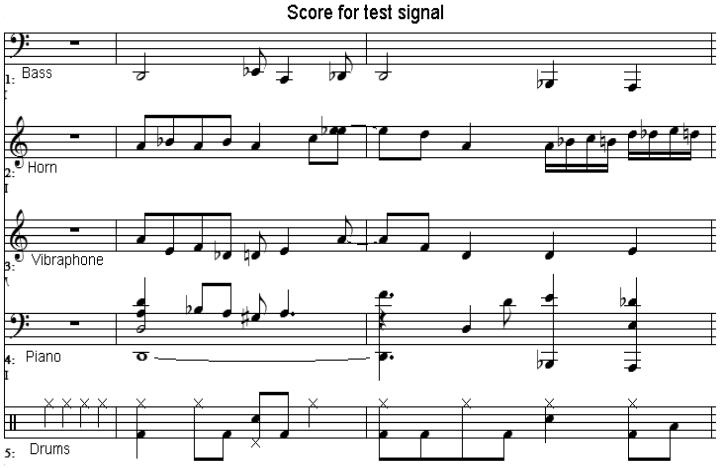

Figure 3.5: The score which was generated for the 5 instruments.

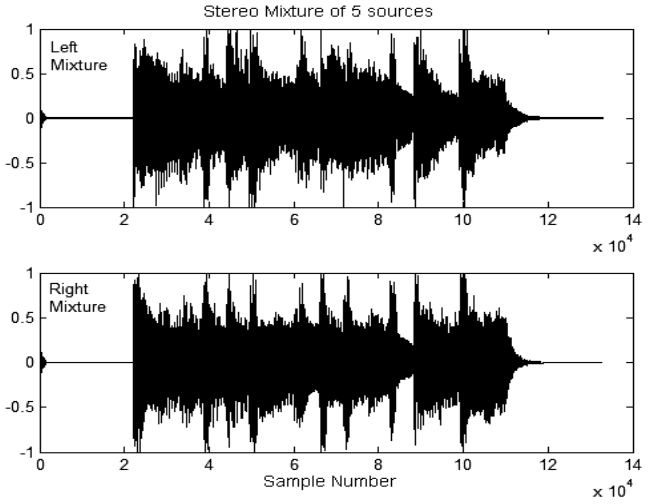

Figure 3.6: The Stereo Mixture containing 5 panned sources.

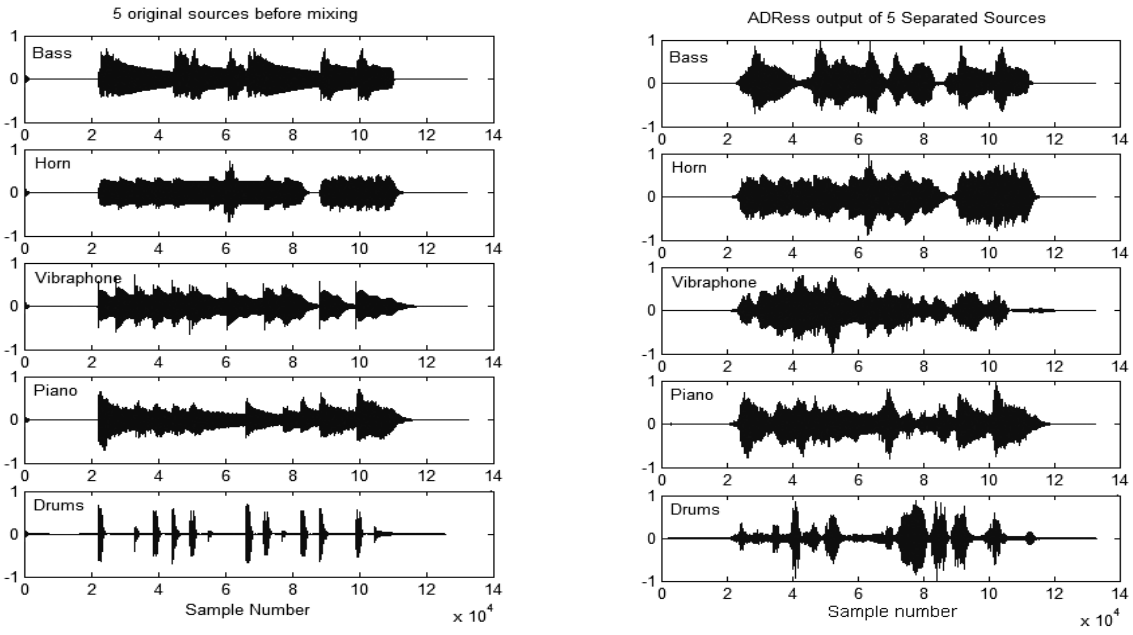

Figure 3.7: The 5 original sources before mixing and after separation.

The resulting separations are of reasonably high quality. There are some obvious visual differences between the input and output time domain plots and there are some obvious audible artefacts but the quality is significantly high. Furthermore when the separations are 'remixed', the resultant mixture is almost free from artifacts. These audio examples and others can be accessed in (Barry 2019).

3.8 - CONCLUSIONS

We have presented an algorithm which is able to perform sound source separation by decomposing stereo recordings into frequency-azimuth subspaces. These subspaces can then be resynthesised individually, resulting in source separation. The only constraints are that the recording is made in the fashion described in Section 2, and

that the sources do not move position within the stereo field. We feel that ADRess is applicable to a large percentage of commercial recordings.

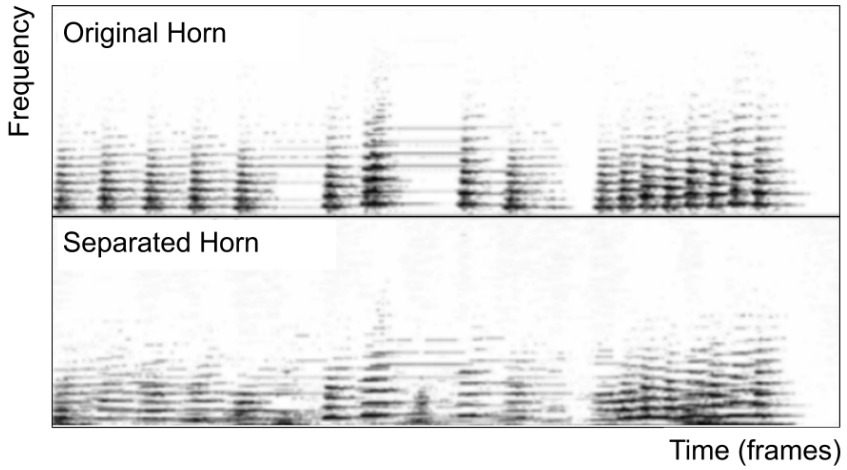

Figure 3.8: The spectrogram here contains the original horn part on top and the separated horn part using ADRess on the bottom.

Figure 3.8 compares the spectrogram of the horn from prior to mixing (top) and after separation (bottom). It should be clear to see that the general features of the horn have been captured well but there are visible and audible artefacts from other sources included in the separation.

3.8.1 - Future Work

It is apparent when listening to the audio separations from ADRess that transients are often smeared. This results from the fact that the window length of 4096 samples is chosen in order to give adequate frequency resolution for lower octaves. However, this window size does not afford adequate time resolution for higher frequencies and transients. A multi-resolution approach could mitigate this problem. By splitting the

audio into 2 or more bands, different window sizes could be applied in order to achieve better transient response without affecting lower frequencies.

Also worthy of further investigation is the possibility of automatically choosing the best algorithm parameters (azimuth and width) based on the observed properties of the frequency azimuth plane.

3.9 - REAL-TIME ADDITIONS

The paper above alluded to the ability to operate in real time but omitted the details of how this could be achieved. The specific real-time buffering scheme used was published later in (Barry et al. 2008). What follows is an excerpt from that article describing the scheme.

Real-time Buffer Scheme

One of the key issues in a real-time implementation is the choice of buffer scheme and for completeness sake we suggest a suitable scheme here. In offline processing, the entire signal is overlapped and concatenated before playback. However, in a real-time environment, a constant stream of processed audio must be outputted and consecutive output frames must be continuous. In order for seamless concatenation, the boundaries of each output frame must be at the constant gain associated with the overlap factor in order to avoid modulation. The method presented below addresses this concern. For reasons discussed in previous sections, a 75% overlap is recommended. This effectively means that at any one time instant, 4 analysis frames are actively contributing to the current output frame.

In Figure 3.9, the audio to be processed is divided into overlapping frames of length *N*. In order to output a processed frame, 4 full frames would need to be processed and overlapped. This leads to considerable latency from the time a parameter change is affected to the time when its effects are audible at the output.

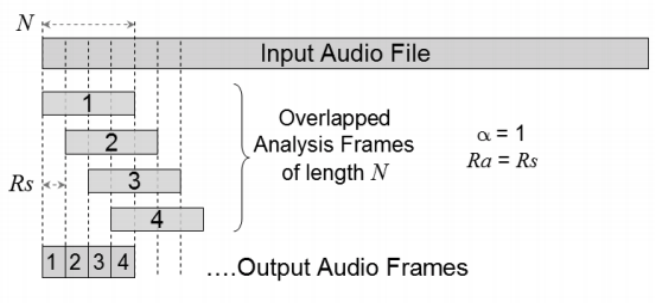

Figure 3.9. The relationship between input and output frames for α =1

However, given that the synthesis hop size is fixed at R_s=R_a, we can load and process a single frame of length N, output ¼ of the frame, and retain the rest in a buffer to overlap with audio in successive output frames. To do this, a buffer of length N is required in which the current processed frame (with synthesis window applied) is placed. Three additional buffers of length $3N/4$, $N/2$ and $N/4$ will also be required to store remaining segments from the three previously processed frames. Each output frame of length $N/4$ is then generated by summing samples from each of these four buffers.

Figure 3.10 shows how the buffer scheme works. On each iteration u, a full frame, F^u, of length N is processed and placed in buffer 1. The remaining samples from the three previous frames occupy buffers 2, 3 and 4. The required output frame of length $N/4$, S^u, is generated as defined in equation 3.11.

$$S^u(n) = F^u(n) + F^{u-1}(n+N/4) + F^{u-2}(n+N/2) + F^{u-3}(n+3N/4)$$
$$\forall n \qquad 1 \le n \le N/4 \tag{3.11}$$

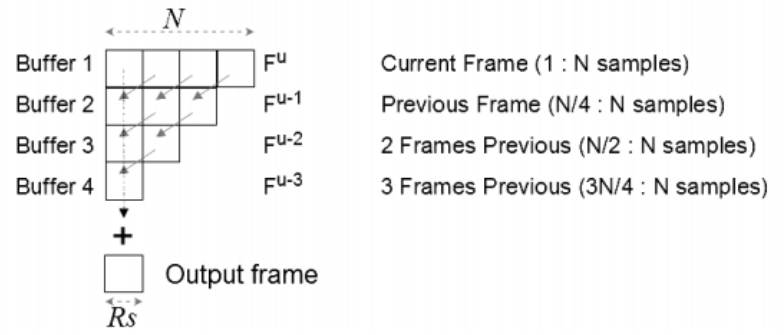

Figure 3.10. Real-time output buffer scheme using a 75% overlap. The gray arrows indicate how each segment of each buffer is shifted after the output frame has been generated.

From equation 3.11, it can be seen that the output frame, $S^u(n)$, is generated by summing the first $N/4$ samples form each buffer. Once the output frame has been generated and outputted, the first $N/4$ samples in each buffer can be discarded. The data in all buffers must now be shifted in order to prepare for the next iteration. The gray arrows in Figure 3.10 illustrate how each segment of each buffer is shifted in order to accommodate a newly processed frame in the next iteration. The order in which the buffers are shifted is vital. Buffer 4 is filled with the remaining $N/4$ samples from buffer 3, buffer 3 is then filled with the remaining $N/2$ samples from buffer 2, and finally buffer 2 is filled with the remaining $3N/4$ samples from buffer 1. Buffer 1 is now empty and ready to receive the next processed frame of length *N*. The result of this scheme, is that ¼ of a processed frame will be outputted at time intervals of R_s, which is equal to $N/4$ samples. Using the suggested frame size of 4096 samples, the output will be updated every 1024 samples which is approximately equal to 23.2 milliseconds. The audio will be processed with newly updated parameters every 23.2 milliseconds, but the latency will be larger than this and depends on the time required

to access and write to hardware buffers in the audio interface. In general, it is possible to achieve latencies < 40ms.

3.10 COMPARATIVE TESTING

A set of objective testing measures to compare source separation algorithms was proposed in (Vincent et al. 2006). The following year, an open blind source separation evaluation campaign was undertaken and the results were published in (Vincent et al. 2007). Within this campaign, ADRess was compared against 10 other algorithms using four objective measures across 4 audio mixtures with various numbers and types of sources. The four objective measures detailed in (Vincent et al. 2006 and 2007) are as follows:

- Source Image to Spatial distortion Ratio (ISR)
- Source to Interference Ratio (SIR)
- Sources to Artifacts Ratio (SAR)
- Source to Distortion Ratio (SDR) which is a weighted average of the previous 3 metrics

The four audio mixtures are described as follows:

- **Female**: Four female voices speaking simultaneously in different languages, panned to different locations in the stereo mixture.
- **Male:** Four male voices speaking simultaneously in different languages, panned to different locations in the stereo mixture.

- **No Drums:** Electric Guitar, Acoustic Guitar and Bass panned to different locations in the stereo mixture.
- **With Drums:** Bass, kick, snare and hi hat panned to different locations in the stereo mixture. For clarity the kick and snare are panned to the same location.

Each of the algorithms tested in the evaluation campaign is summarised in Table 3.1 below.

No.	Submiter Name	Source Localisation Method	Source Estimation Method
1	D. Barry (ADRess)	Manual IID clustering from a magnitude-weighted histogram with auditory feedback (Barry et al, 2004)	Source magnitude estimation in the STFT bins associated with each IID cluster (Barry et al, 2004)
2	P. Bofill	Peak picking on a smoothed IID histogram with STFT bins selected as in (Xiao et al, 2005)	Minimization of the L1 norm of the real and imaginary parts of the source STFTs (Bofill et al, 2006)
3	A. Ehmann	Manual peak picking on an IID histogram	Binary STFT masking with different resolutions at high/low frequencies
4	V. Gowreesunker	Peak picking on a thresholded IID histogram (Gowreesunker et al, 2007)	Binwise MDCT projection onto the nearest IID subspace (Gowreesunker et al, 2007)
5	M. Kleffner	Peak picking on a thresholded IID histogram (Mohan et al, 2003) with STFT bins selected as in (Arberet et al, 2006)	Online FFT-domain minimum variance beamforming (Lockwood et al, 2004)
6	N. Mitianoudis	Soft IID clustering given the number of sources (Mitianoudis et al, 2007)	Binwise MDCT projection onto the nearest IID subspace (Mitianoudis et al, 2007)
7	H. Sawada	Hard IID clustering given the number of sources (Sawada et al, 2007)	Binary STFT masking
8	E. Vincent	Manual peak picking on an IID histogram weighted as in (Arberet et al, 2006)	Minimization of the I0 norm of the source STFTs (Vincent 2007)
9	M. Xiao	Hard fixed-width IID clustering on selected STFT bins (Xiao et al, 2005)	Mixing inversion with 2 sources per time frame estimated from the mixture covariance (Xiao et al, 2005b)
10	M. Xiao	Hard fixed-width IID clustering on selected STFT bins (Xiao et al, 2005)	Extension of (Xiao et al, 2005b) with more active sources in some time frames
11	R. Weiss & M.Mandel	Soft (IID,IPD) clustering given the number of sources (Mandel et al, 2007)	Soft STFT masking by cluster probabilities (Mandel et al, 2007)

Table 3.1 - Summary of algorithms compared in the 2007 BSS Evaluation Campaign. Reproduced from (Vincent et al. 2007)

The original published results ranked ADRess in 6th place out of 11 algorithms with respect to the overall SDR measure. However, the authors noted that the submitted test results from ADRess were hampered due to the fact that they contained *"strong time-localized interference within the last 100 ms of each estimated source image signal"* (Vincent et al. 2007 b). This led to degradation of the objective results for the

ADRess algorithm, the cause of which was an error on my part when processing the test material. After the fact, the ADRess algorithm was retested with the processing error corrected. Table 3.2 shows the results as published in (Vincent et al. 2007) but with the retested results for ADRess once the error had been corrected. The results show that ADRess ranked number 2 in 11 out of 14 examples for the SDR measurement. ADRess was outperformed by Vincent's own algorithm (Vincent 2007) in 12 of the examples. However, ADRess did rank first for average ISR performance. Table 3.2 below uses a colour key to indicate how ADRess ranked for SDR in each individual example. Although it was not the focus of the evaluation campaign, the authors noted that ADRess was the only algorithm operating in real time or faster. Figure 3.11 and 3.12 depict waveform comparisons between separations produced by ADRess and (Vincent 2007) for all 14 sources across all 4 mixtures. Column A shows all of the original mixtures prior to mixing, column B shows all the separations from ADRess and column C shows all the separations from (Vincent 2007). From inspection, it can be seen that the separations produced by ADRess and (Vincent 2007) are broadly similar in a visual sense. Generally, they tend to succeed and fail in capturing the original waveforms on the same examples. For example, in *"No Drums 1 - Bass"* in Figure 3.12, it can be seen that both ADRess and (Vincent 2007) captured the waveform characteristics well but in *"No Drums 2 - Electric Guitar",* both algorithms failed to capture the waveform in a visual sense. However, there are some examples where one captures the original waveform better than the other. In *"With Drums - Hi hat"* in Figure 3.12, ADRess appears to capture the waveform characteristics better than (Vincent 2007).

		Female				Male				No Drums			With Drums		
	(dB)	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T1	T2	T3
Algorithm 1	SDR	11.3	6.4	5.5	7.8	13.0	2.7	5.8	6.6	16.2	1.3	6.6	7.7	4.9	22.7
(retested)	ISR	21.3	19.0	12.0	12.9	24.0	8.2	11.9	11.2	22.8	12.2	10.2	16.9	7.8	41.9
D. Barry	SIR	19.6	12.3	11.2	17.8	21.8	9.3	13.3	19.3	23.5	6.1	9.7	14.0	17.3	25.5
(1 s/mix)	SAR	11.8	7.3	5.7	8.3	13.6	1.6	5.8	6.1	18.3	2.1	9.9	9.5	4.5	26.1
Algorithm 2	SDR	3.6	6.2	4	3.4	3.8	1.5	4.8	3.2	8.3	-3.2	6.3	3.1	5.7	6.9
P. Bofill	ISR	3.8	10.3	11.6	6.1	3.9	8.3	11.7	6.7	8.6	11.3	8.3	8.3	8.4	7
(5 min/mix)	SIR	18	10.6	7.8	12.6	20.6	2.6	9.6	11.2	20	-2.9	15	4.6	14.2	33.8
	SAR	13.5	7.6	6	9.7	16.2	2.7	6.5	8.6	20.1	4.8	14.1	6.3	5.8	25.2
Algorithm 3	SDR	10.3	5.9	4.6	7.5	11.1	1.3	4.1	6.6	15.1	-6.8	3.7	7.2	3	20.5
A. Ehmann	ISR	17.1	12.6	11.9	11.7	18.3	8.9	14.5	11.9	20	8.9	5.6	15.8	4.7	31.1
(5 s/mix)	SIR	22.3	14	9.4	19.9	23.9	6.2	7.6	17.3	25.8	-4.9	18.1	14.2	15.1	27.7
	SAR	10.5	6	5.2	8	11.5	0.7	5.9	6.3	17.3	-0.6	3.6	8.9	2.1	21.8
Algorithm 4	SDR	2.7	4.3	3.1	3.8	2.8	0.5	3.9	3.4	5.7	-2.2	3.4	4.8	3.3	8
V. Gowreesunker	ISR	2.9	7.3	7.4	6.1	2.9	5.7	8	6.1	6.1	8	5	6.2	4.8	10
& A. Tewfik	SIR	18.9	9.7	7.7	15.9	20.2	2.1	10.4	15	27.1	-2	18.6	12.1	17.2	25.1
(10 s/mix)	SAR	8.5	4.3	2.7	5.7	9	-0.4	3.8	4.2	10.9	1.3	4	6.8	2.2	12.3
Algorithm 5	SDR	-20.5	-26.8	-22	-21.8	-19.2	-29.6	-26.6	-20.5						
M. Kleffner	ISR	-19.2	-25.7	-20.4	-20.6	-18.1	-26.6	-25.2	-18.7						
(10 min/mix)	SIR	18.6	12.6	9.2	14.3	19.9	4.6	10.9	11.9						
	SAR	5.8	7.2	6.2	6.5	6.6	3.6	6.3	4.9						
Algorithm 6	SDR	-17.2	-21.4	-18.9	-18.3	-17.8	-23.5	-17.9	-17.4	-11.1	-25.1	-10.2	-12.1	-10.9	-8.5
N. Mitianoudis	ISR	-15.9	-18.7	-14.7	-16.4	-17	-17.4	-14.8	-15.2	-9.8	-13.5	-6.2	-10.6	-7.6	-8.4
(3 min/mix)	SIR	19	6.3	7.9	16.9	22.5	2	10.8	16	13.3	-7.5	17.5	16.6	15.9	26.7
	SAR	6.1	4.3	0.5	4.1	8.1	-0.4	2.1	3.5	8.5	2.5	1.4	6.5	2.7	20.1
Algorithm 7	SDR	9	5.6	4	6.2	10.8	0.5	5	6	15.5	-0.8	4.7	9.1	5.1	17.5
H. Sawada	ISR	15	14.6	11.7	9	17.3	9.9	13.1	9.8	20.9	15.6	6.9	16.9	8.5	30.7
(9 s/mix)	SIR	20.8	11.2	7.7	20.3	22.7	4	9.5	17.5	23.5	2.8	18.3	17.1	18	25.7
	SAR	9.1	6.3	4.8	7	11.2	1.1	5.8	5.7	17.8	3.1	5.2	10.5	4.5	18.3
Algorithm 8	SDR	13.3	6.8	5.8	8.4	16.4	3.2	6.1	7.9	22.2	2.7	16.8	-0.6	3.1	28.3
E. Vincent	ISR	26.6	11.1	11.4	18.2	28.3	7.4	10.9	20	32.1	10.2	27.6	8.8	4.5	46.3
(5 s/mix)	SIR	17.4	16.8	12	13	20.4	8.4	13.9	12	27.8	7.8	22	1.9	17	29.8
	SAR	15.4	7.1	6.1	10	18.8	2.5	6.3	9.6	24.1	3.2	18.5	7.7	2.5	34.1
Algorithm 9	SDR	9.1	2.1	0.4	3.7	12.6	-0.8	1.6	3.6	14	-5.3	8.8	-0.7	3	26
M. Xiao	ISR	23.1	8.3	8.7	10.3	27.6	4.2	9.6	12.6	29.8	1.4	19.8	8.9	8.3	43.9
(2 s/mix)	SIR	14.7	9.6	3.6	10.9	17.6	3.1	6.2	9.6	19.3	-7.1	14	2.2	9.9	33
	SAR	10.4	0.8	0.6	3.1	14.2	-4.2	0.9	3.7	15.6	-7.3	9.7	1.1	1.4	27.1
Algorithm 10	SDR	8.1	-1.5	-2.1	2.5	12.3	-8	-0.5	2.3	14.2	-13.6	13.5	-2.8	3.6	8.1
M. Xiao	ISR	27.2	18	18.6	20.7	29.4	15	21.4	23.1	29.9	5.8	32.3	6.4	20.9	14.3
(2 s/mix)	SIR	11.1	3.2	-0.5	5.4	15.2	-3.5	1.9	4.4	19.3	-10.6	14.2	-0.4	5.6	25.9
	SAR	11.1	1.5	5	6.1	15.5	-1.1	5.4	7.7	15.8	3.4	22.2	8.2	8.4	8
Algorithm 11	SDR	3.6	0.3	-3.4	0.2	1.5	-8.5	0.4	0.3	4	-9.2	-2	-11.8	-9.8	4
R. Weiss	ISR	5.5	0.3	12.4	0.2	1.7	10.1	0.6	0.5	4.5	5	5.8	0	6.2	4.4
and M. Mandel	SIR	5.2	-5.7	-4.7	-7.2	3.1	-9.9	-5.6	-4.8	15.8	-13.6	-3	-18.8	-10.7	16.8
(20 min/mix)	SAR	5.6	0.1	9.9	-0.7	3.2	11.7	1.5	-0.6	6.5	8.4	0.1	2.4	3.5	7.1

Table 3.2 - Blind Source Separation Evaluation Campaign results (Vincent et al. 2007) with retested ADRess results. Ranking key: Red = 1 , Yellow = 2, Blue = 3, Green = 4

*Figure 3.11 - Waveform comparisons for speech separation between **A**: Original sources prior to mixing, **B**: ADRess separations and **C**: (Vincent 2007) separations*

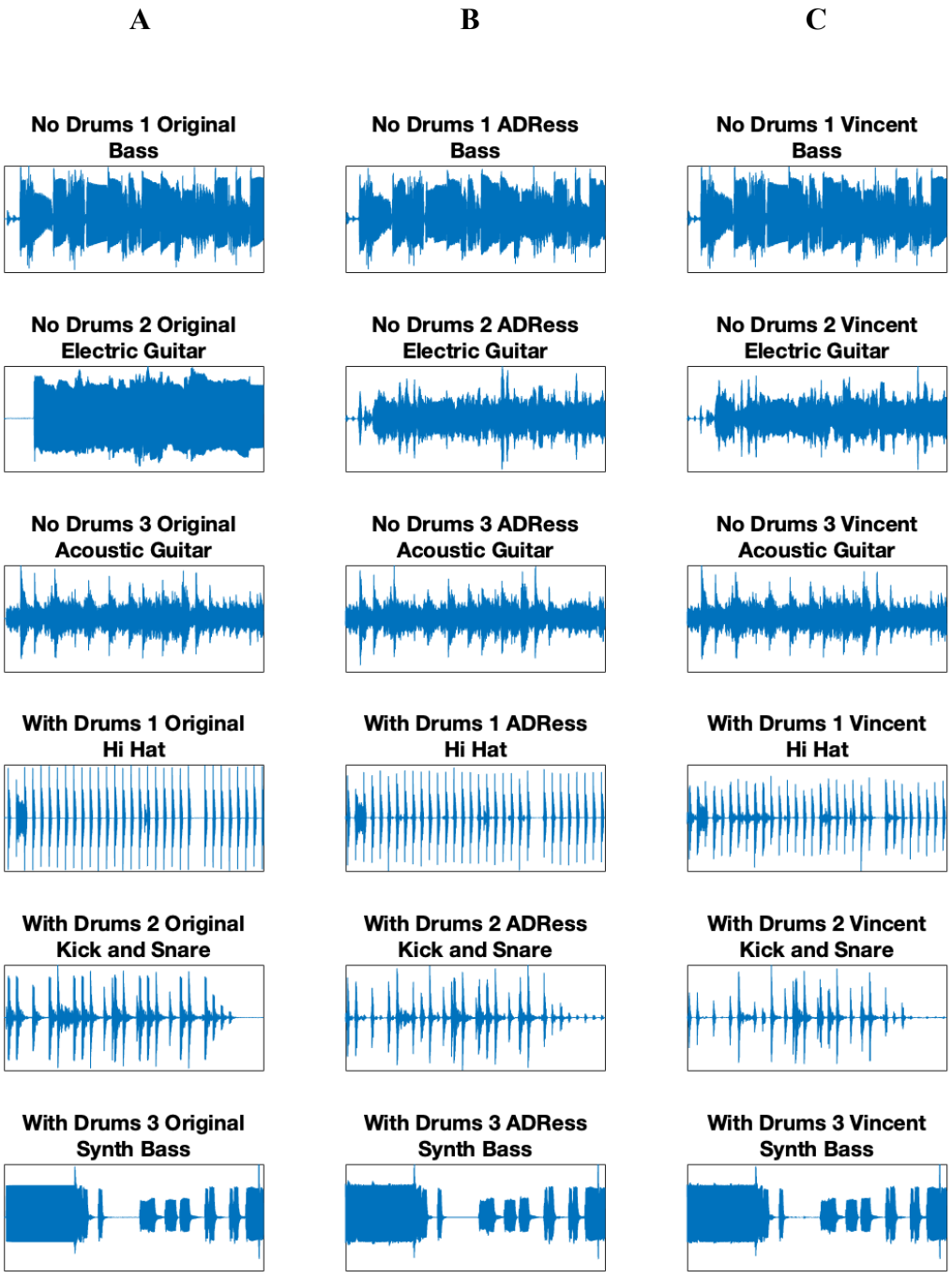

*Figure 3.12 - Waveform comparisons for music separations between **A**: Original sources prior to mixing, **B**: ADRess separations and **C**: (Vincent 2007) separations*

Audio examples for all of the test material described above can be accessed at (Barry 2019).

3.10.1 Subjective Audio Quality

Subjective testing was not conducted as part of the open evaluation campaign described in the previous section but I would encourage the reader to compare the audio results available at the website accompanying this document (Barry 2019). Here, I would like to describe the subjective quality of separated sources using the ADRess algorithm. Firstly, it should be noted that both the subjective and objective quality achievable will depend greatly on the following characteristics of the mixture:

1. The number of sources in the mixture. The more sources present, the worse any single separated source is likely to be in terms of subjective audio quality.
2. The pan position of the sources in the mixture. If sources are panned to the same position in the mix, they cannot be separated. Unique and maximally distant pan positions will lead to greater flexibility for separation
3. The amount of time-frequency overlap between the sources. If sources occupy the same time-frequency bins in the STFT, sub-optimal separation will result. W-DO sources will separate optimally given that they do not overlap significantly in a time-frequency representation.
4. The algorithm parameter settings. Minimising neighbouring source interference through narrow azimuth subspace width settings will usually lead to more separation but at the cost of timbral fidelity.

The limitations described above lead to the following subjective artefacts in the separated sources.

Interference from neighbouring sources: The most common artefact associated with the source separation process is interference from neighbouring sources. This interference can range from clearly audible traces of other sources down to barely noticeable residual noise which manifests as gurgle or bell type sounds. In general, smaller values for the azimuth subspace width will lead to more gurgle/bell artefacts but with greater suppression of neighbouring sources. Larger values, will lead to a higher fidelity reproduction of a single source but with very audible traces of neighbouring sources.

Phasiness artefacts: Phasiness is normally associated with the phase vocoder but is common in many processes which modify either the magnitude spectrum or phase spectrum. It is sometimes described as sounding like reverberation or a swishing sound in the background. It is generally caused by the fact that the phase values are no longer valid for the modified magnitude spectrum (or vice versa) during the inversion process. In the case of ADRess, only the magnitude spectrum is modified and the original phases are used. As a result, the separated source is forced to use the values obtained from the mixture of all sources. This is intuitively suboptimal and leads to the phasey artefacts described.

Transient Smearing: In order to achieve good frequency resolution in lower octaves, we use an analysis window size of 4096 samples at 44100 Hz. This gives an approximate bin width of 10.71 Hz which is required for separation of low bass notes. Although it gives a suitable frequency resolution, this window size is too large to give good temporal resolution. Because of this, rapidly changing temporal events such as

transients are poorly represented in the time-frequency domain and when subjected to the source separation process are often badly corrupted. Effectively, the same transient ends up being processed slightly differently in four consecutive frames (due to 75% overlap) which results in transient smearing upon resynthesis. Timbrally, it sounds as if the sharp attacks of transients have been softened. The degree to which this happens depends largely on the four factors listed above.

Despite the artefacts described above, the subjective quality of ADRess is adequate for many of its uses including music education and upmixing. In chapter 6, we illustrate that these artefacts can be masked when ADRess is used for the task of upmixing.

In this chapter my principal novel contribution has been presented - the ADRess algorithm. The following four chapters present further contributions built upon the ADRess algorithm. The next chapter explores two alternative signal reconstruction methods for the ADRess algorithm which ordinarily uses an inverse fast fourier transform to synthesise the separated source(s). The two alternative algorithms explored in the following chapter are "*Magnitude Only Reconstruction*" and "*Sinusoidal Modeling*". The former was chosen because the ADRess algorithm makes no attempt to do phase reconstruction, only magnitude reconstruction. The latter was chosen to mitigate some of the frequency domain artefacts which can be present in the ADRess outputs.

CHAPTER 4: COMPARISON OF SIGNAL RECONSTRUCTION METHODS FOR THE AZIMUTH DISCRIMINATION AND RESYNTHESIS ALGORITHM

This chapter presents a minor contribution of this dissertation which extends my work on the Azimuth Discrimination and Resynthesis algorithm (ADRess). It was originally published at the Audio Engineering Society Convention in 2005 (AES 118) and is presented here in its entirety. The paper included co-authors Eugene Coyle and Bob Lawlor who acted as my PhD supervisors at the time.

4.1 - ABSTRACT

The Azimuth Discrimination and Resynthesis algorithm, (ADRess), has been shown to produce high quality sound source separation results for intensity panned stereo recordings. There are however, artifacts such as phasiness which become apparent in the separated signals under certain conditions. This is largely due to the fact that only the magnitude spectra for the separated sources are estimated. Each source is then resynthesised using the phase information obtained from the original mixture. This paper describes the nature and origin of the associated artifacts and proposes alternative techniques for resynthesising the separated signals. A comparison of each technique is then presented

4.2 - INTRODUCTION

The ADRess algorithm (Barry et al. 2004) and (Barry et al. 2004 b) performs the task of source separation based on the lateral displacement of a source within the stereo field. The algorithm exploits the use of the "pan pot" as a means to achieve image localisation within stereophonic recordings. As such, only an interaural intensity difference exists between left and right channels for a single source. Gain scaling and phase cancellation techniques are used in the frequency domain to expose frequency dependent nulls across the azimuth plane. The position of these nulls in conjunction with magnitude estimation and grouping techniques are then used to estimate the spectra of the separated sources.

Although the magnitude spectra are good approximations of the original source spectra, the algorithm makes no attempt at finding a set of phase approximations for source resynthesis. Instead, the phase information taken from the original mixture is used for all sources. This is shown to be acceptable in the majority of cases but artifacts such as phasiness can exist. This is particularly noticeable in percussive or transient audio. Other artifacts can arise when two sources overlapping in the time-frequency domain are positioned in close proximity to each other in stereo space. These artifacts are the result of what is identified as 'frequency-azimuth smearing' in (Barry et al. 2004 b). Effectively, low energy sources can be significantly degraded by high energy sources in the stereo mixture. For example, a sustained note within one separation may contain amplitude modulation or even complete dropouts due to the onset of a drum which has been panned to a similar position.

The signal reconstruction in the original ADRess algorithm is achieved by inverting the short-time Fourier Transform (STFT) of the separated source spectra with the original mixture phases. In this paper we explore the use of alternate signal reconstruction methods. Since there is no method for determining the original phase contributions of each source in a mixture, we must rely solely on the magnitude spectra of the separated sources. For this reason, the "magnitude-only" reconstruction technique in (Griffin et al. 1984) is proposed. A Sinusoidal Model (McAuley et al. 1986) resynthesis is also presented here as an alternative reconstruction method.

The separated spectra produced by ADRess are simply estimates of the actual source spectra and as such may be distorted, i.e. the lobes associated with peaks in the frequency domain can become smeared which would lead to artifacts on resynthesis. A sinusoidal model reconstruction may provide better results on the basis that only the peaks in the frequency domain are extracted for resynthesis.

4.3 - BACKGROUND

The ADRess algorithm achieves source separation by taking advantage of destructive phase cancellation in the frequency domain. One channel is iteratively gain scaled and subtracted from the other in the complex frequency domain after which the modulus is taken. The resulting array is of dimension $N \times \beta$, where N is the number of frequency points and β, the azimuth resolution, is the number of equally spaced gain scalars between 0 and 1. The operation reveals local minima, due to phase cancellation, across the azimuth plane for each frequency component. Components belonging to a single source are seen to have their minima in a localised region about some gain scalar which ultimately refers to the pan position of the source in stereo space.

The process can be described as follows; firstly we take the fast Fourier transform (FFT) of a windowed (typically raised cosine) short time segment of length N of each channel,

$$Lf_{(k)} = \sum_{n=0}^{N-1} L_{(n)} W_n^{kn} \qquad (4.1)$$

where $W_n = e^{-j2\pi/N}$ and similarly for the right channel yielding *Lf(k)* and *Rf(k)* which represent short time complex frequency representations of the left and right signal. The iterative gain scaling process results in what is termed a 'frequency-azimuth plane' and is constructed using equation 4.2,

$$AzL_{(k,\,i)} = \left| Rf_{(k)} - g_{(i)}.Lf_{(k)} \right| \qquad (4.2)$$

where $1 \leq k \leq N$ and where $g(i)=i/\beta$, for all i where, $0 \leq i \leq \beta$, and where i and β are integer values.

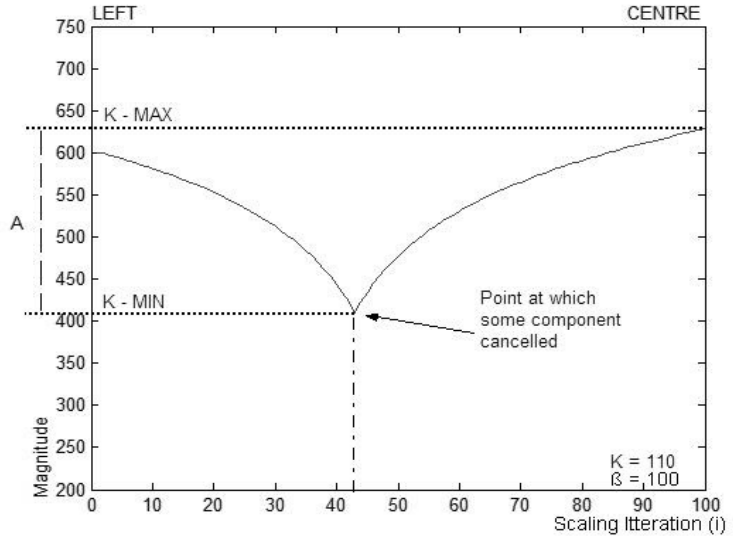

Figure 4.1: One channel is iteratively gain scaled and subtracted from the other in the complex frequency domain for each bin. A local minimum in this function occurs at the point of maximum phase cancelation. This point is deemed to be the azimuth location of that frequency component.

β refers to the number of gain scalars to be used and ultimately gives rise to the resolution achieved in the azimuth plane. For example, β=10, will result in 10 discrete azimuth positions for each channel, i.e. 20 positions from left to right. equation 4.2 represents the left half of the azimuth plane, *AzL(k,i)*; the right half is created by changing the positions of the left and right variables above. figure 4.1 shows the result of the above function for one frequency component, k=110.

In figure 4.2, it can be seen that the minima for multiple components from two sources align along the relevant source positions. These local minima represent the points at which frequency components experience a drop in energy due to destructive phase cancellation. This energy drop is directly proportional to the amount of energy which the cancelled source had contributed to the overall mixture and so to invert these minima around a single azimuth point should yield short-time magnitude spectra of the individual sources. To do this inversion we simply subtract the minimum from the maximum of the function as shown in figure 4.1 and described by equation 4.3.

To invert the minima we use equation 4.3.

$$AzR(k,i) = \begin{cases} AzR(k)_{\max} - AzR(k)_{\min} & \text{if } AzR(k,i) = AzR(k)_{\min} \\ 0, \text{ otherwise} \end{cases} \quad (4.3)$$

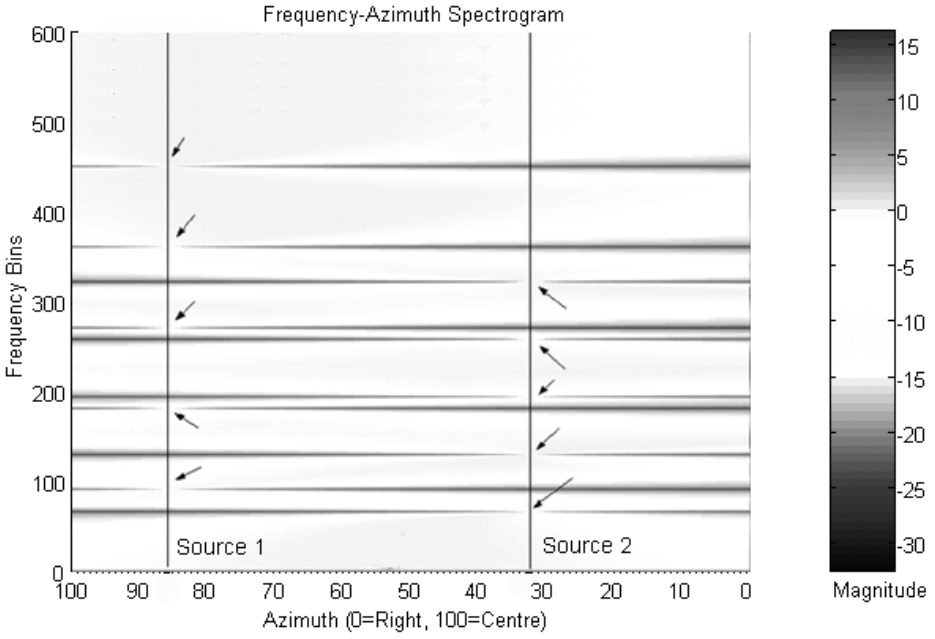

Figure 4.2: Local minima for 2 complex sources.

The effect of this operation is to turn the minima or nulls into peaks. equation 4.3 must be performed for both left and right frequency azimuth planes. At this point we have separated out all frequency components according to the azimuth positions at which they cancelled. It is the case that frequency components and their relative magnitudes relating to a single source will be grouped around a single azimuth position which corresponds to the pan position of the source. In order to resynthesise a source, we simply extract the portion of the frequency azimuth plane around an azimuth position using equation 4.4,

$$YR_{(k)} = \sum_{i=d-H/2}^{i=d+H/2} AzR(k,i) \qquad \begin{matrix} 1 \le k \le N \\ 0 \le d \le \beta \end{matrix} \qquad (4.4)$$

where *d* is the azimuth index, i.e. the azimuth position of the source for separation and *H* is the azimuth subspace width which is simply a neighborhood around the azimuth index. *YR(k)* is now an *N x 1* array containing the short-time magnitude spectrum of a single source or azimuth subspace. Typically at this point, we use an IFFT with the original mixture phases and a standard overlap add technique to resynthesise the signal. One problem is that the estimated spectra no longer have the windowed characteristics of the signal due to the ADRess process. For this reason a synthesis window must also be applied to avoid discontinuities in the resynthesised signal. Furthermore, the overlap is set at 3/4 the frame size (75%) to avoid modulation in the resynthesis since we have effectively windowed the data twice. This reconstruction method gives satisfactory results even though no phase estimates are provided for the separated sources. In the next section, we attempt a reconstruction with only the magnitude spectra which ADRess produces.

4.4 - MAGNITUDE ONLY RECONSTRUCTION

In (Griffin et al. 1984), the authors propose an iterative technique which allows a signal to be reconstructed, given only the modified short-time Fourier transform magnitudes (MSTFTM) and a set of initial, or even random phases. The approach is based on the fact that not all STFTs are ‘valid’ in the sense that there may not exist a sequence of time values which would yield a given STFT. This is the case with many frequency-domain techniques for sound source separation, in that, typically only the

magnitude spectra of the sources are estimated. These estimated spectra do not correspond to any 'real' signal. The algorithm in (Griffin et al. 1984) attempts to find a real signal whose STFT is closest in a least squared error sense to the MSTFTM which is provided. Using a standard windowed overlap add procedure, the algorithm iterates between the time and frequency domain. During each iteration the phases are altered due to the influence of two consecutive frames overlapping, however, the re-synthesis for any given iteration always uses the original MSTFTM and the updated phases. It is shown by the distance measure described by equation 4.5, that the squared error between the STFT of the real signal and the MSTFTM is reduced in each iteration. Through this process a set of phase approximations can be arrived at. As the iterations increase, the phase estimates become more accurate until a critical point is reached, after which no significant improvement is achieved.

$$D_i[x^i(n), y_w(mS,\omega)] = \sum_{m=-\infty}^{\infty} \sum_{\omega=-\infty}^{\infty} [\left|x_w^i(mS,\omega)\right| - \left|y_w(mS,\omega)\right|]^2 \tag{4.5}$$

D_i represents the distance between the STFT of the resynthesised signal after the i^{th} iteration, $|x_w^i(mS,\omega)|$, and the given MSTFTM, $|y_w(mS,\omega)|$, where m is a frame index and S is the hopsize. In equation 4.5, $x^i(n)$, is notated as such to emphasize the fact that $x_w^i(mS,\omega)$ is a valid STFT, whereas $y_w(mS,\omega)$ may not be. For the i^{th} iteration then, the resynthesised signal is given by equation 4.6.

$$x^i(n) = \sum_{\omega=-\infty}^{\infty} \left|y_w(mS,\omega)\right| \cdot e^{j\angle x_w^{i-1}(mS,\omega)}, i > 1 \tag{4.6}$$

For the first iteration, $i=1$, a set of random phases are chosen. The purpose of using this algorithm as a resynthesis method for ADRess was to determine whether a better set of phase approximations could be arrived at than simply using the original mixture phases. The distance measure D_i , given by equation 4.5, was used to ascertain which set of phase estimates give the best resynthesis in a least squared error sense. Furthermore, the original mixture phases were used as the initial phase estimates for a magnitude only reconstruction to see if the algorithm would converge to even better phase estimates with fewer iterations. Figure 4.3 shows that the distance is reduced for each iteration where the initial phase estimates are random, but the error is never less than that of simply using the original phases, even after 100 iterations.

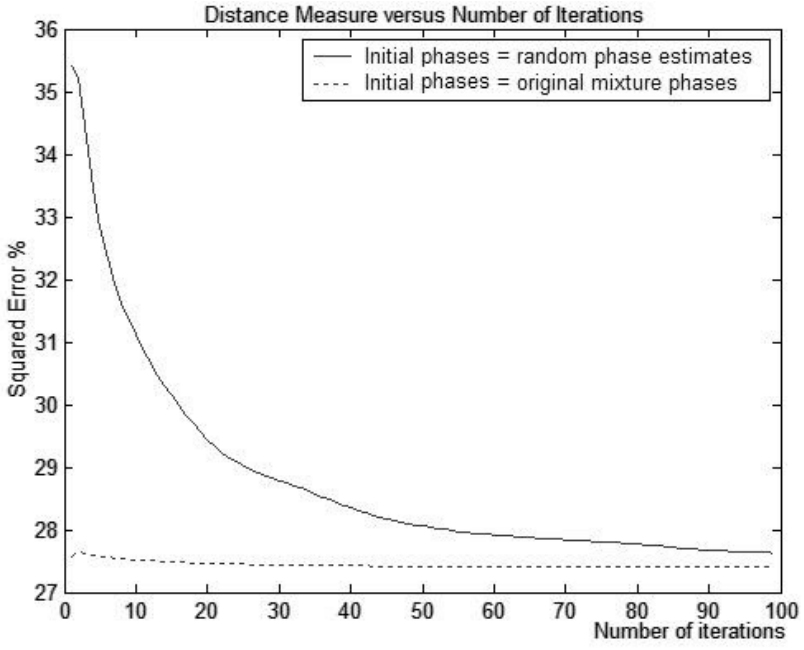

Figure 4.3: The error reduction as a result of several iterations. Note that the iterative phase estimates never improve on the original mixture phase estimates.

Informal listening tests suggest that there is no perceivable advantage to using a magnitude only reconstruction and that the original mixture phases provide better results without any iteration than a magnitude only reconstruction with several iterations. An improved version of the above technique was employed by Slaney for correlogram inversion (Slaney et al. 1994). The principal difference here is that a synchronized overlap-add procedure (Roucos et al. 1985) is used to obtain the optimal frame overlap position to ensure horizontal phase coherence. Ultimately this procedure causes the algorithm to converge with fewer iterations but no perceptual improvement is achieved.

4.5 - SINUSOIDAL MODEL RECONSTRUCTION

Sinusoidal modeling is a well known analysis/synthesis technique for sound modeling and manipulation (McAuley et al. 1986) and (Serra, 1997). The technique is based on the fact that complex musical signals can be represented as a sum of sinusoids with time varying amplitudes, phases and frequencies. These parameters are generally extracted from a time-frequency representation such as the STFT where a sinusoid is represented by a well defined peak with a predictable lobe according to the windowing parameters used in the analysis stage. A peak is usually regarded as any bin with a magnitude greater than that of its two nearest neighbors. The true frequency of the peak can be calculated using either the phase derivative or by using parabolic interpolation. The magnitude is then taken to be the true maximum of the interpolated curve. A peak continuation algorithm tracks peaks from frame to frame to

form trajectories. It attempts to find a peak in the next frame with a similar amplitude and frequency to a peak in the previous frame within some threshold of frequency deviation. These frequency, amplitude and phase values are then interpolated to create sinusoidal tracks with time varying amplitudes and frequencies which can easily be synthesized. This is referred to as the deterministic synthesis which corresponds to the steady state harmonic portions of a signal. The deterministic signal can be accurately modeled using only the frequency and amplitude parameters of the interpolated tracks. The 'noise like' or stochastic parts of the signal can be estimated by subtracting the deterministic signal from the original signal. In this case however, the deterministic synthesis must contain the instantaneous phase values obtained in the analysis stage. The residual which is assumed to be stochastic, is then usually modeled as time varying filtered noise. The basic sinusoidal model architecture has been described here but there are many heuristics which control the behavior of the peak continuation algorithm. One such heuristic gives us the ability to discard sinusoidal tracks which are shorter than a specified duration. This is of particular interest to us since the separations achieved with the ADRess algorithm are subject to brief interference from neighboring sources. This sort of interference as well as noise, appears as 'speckling' on the spectrogram of the separated source. The ability to remove trajectories with such short duration should allow a cleaner resynthesis of the deterministic parts of the signal. Here we use a modified sinusoidal model implemented by Ellis (Ellis, 2003) to carry out the resynthesis of the separated source spectra generated by the ADRess algorithm. The sinusoidal modeling technique is

quite flexible, but this flexibility comes at a cost; adjusting the algorithm parameters for optimal performance depends largely on the signal characteristics and so configuring the algorithm can be quite tedious. For the example shown in Figure 4.4, the algorithm was configured in such a way as to reject as much noise and neighboring source interference as possible. Trajectories with durations less than 6 frames were also discarded. The source in this case was a saxophone which has been separated from a mixture of piano, bass, saxophone and drums. The sinusoidal model resynthesis although cleaner in the pitched regions suffers from artifacts when parameters are incorrectly set. The task of determining how much of the residual signal belongs to the signal and how much is unwanted noise can be difficult, making threshold setting very much a trial and error procedure. However, the results are compelling, and the sinusoidal model could be adapted for the purposes of an offline resynthesis.

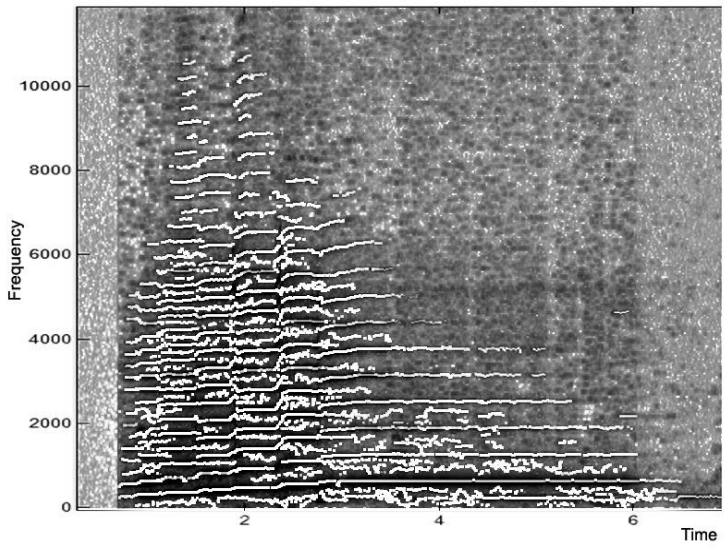

Figure 4.4: Trajectories (shown in white) formed by the peak continuation algorithm superimposed over the spectrogram returned by the ADRess algorithm.

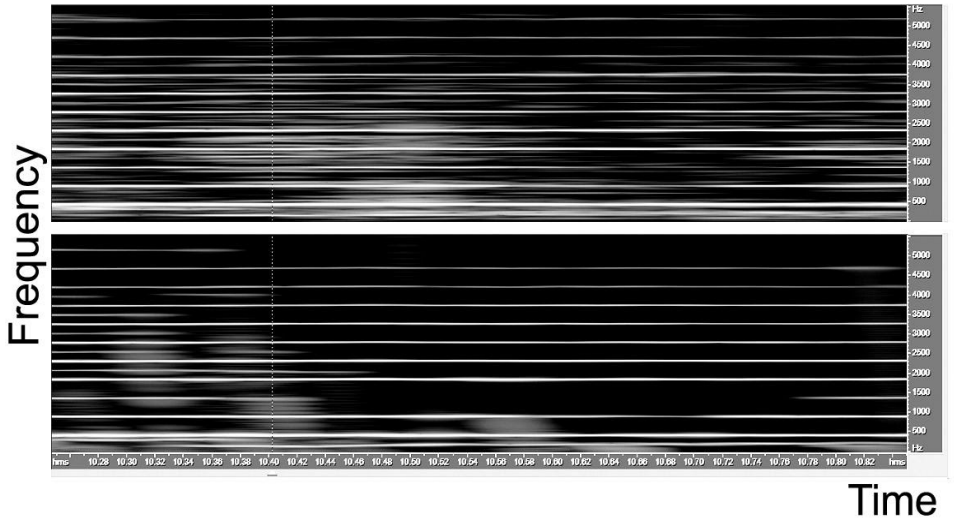

Figure 4.5: Close up on the spectrogram of a pitched region of the saxophone separation with the standard iSTFT method shown on top and the sinusoidal model on bottom.

Figure 4.5 compares the spectrograms of a separated saxophone resynthesised with the standard iSTFT method (top) and the sinusoidal modelling method (bottom). Visibly the the sinusoidal modelling method looks cleaner but it requires significant manual parameter experimentation to make it sound natural.

4.6 - CONCLUSIONS

We have explored the use of two alternative reconstruction techniques for the ADRess algorithm. Firstly the magnitude only reconstruction technique was applied to the separation spectra produced by ADRess in an attempt to arrive at a set of suitable phase estimates. Although the error is reduced significantly after 50 iterations or so using random phase estimates, the error between the initial spectrogram and the final spectrogram is never less than that when the original mixture phases are used. We

believe that the reason for this is linked to a condition identified by (Rickard et al. 2002) known as W-disjoint orthogonality; two sources are said to be W-disjoint orthogonal if there is no significant overlap between the sources time-frequency representations. In the case of musical signals there is usually quite significant overlap in frequency and time, this overlap is the cause of what is identified as 'frequency-azimuth smearing' in (Barry et al. 2004). Effectively when multiple sources contribute to a single frequency component, their phase contributions cause phase cancellation errors in the ADRess algorithm; this in turn causes the frequency dependent nulls to drift away from the apparent azimuth position of a particular source. Sources with the highest intensity will have the most influence over the resultant phases when sources are mixed, and as such will be separated better by ADRess. Furthermore, the phases for any time-frequency point of a mixture of sources will be closest to the phase of the source with the greatest magnitude at that time- frequency point. This leads us to the assumption that there is a variable W-disjoint orthogonality associated with musical mixtures which is purely dependent on the mixture at any given point in time. So for points in time where the sources do not overlap significantly in the frequency domain, the original mixture phases are a close approximation to the source phases.

A sinusoidal model was also applied as a resynthesis technique for the separated source spectra. The technique does offer some advantages for the synthesis of deterministic signals in that some noise and source interference can be rejected resulting in cleaner resynthesis of pitched regions of the signal. The primary

disadvantage is that the technique requires that the operational parameters of the algorithm need to be adjusted accordingly depending on the signal.

The ADRess algorithm has been implemented to run in real time and so computational efficiency is particularly important. Although the reconstruction methods explored here are useful, the method of using the original mixture phases with a standard inverse STFT is still the preferred option as it gives the best trade-off between quality and efficiency.

4.6.1 - Future Work

The main issue associated with resynthesising a modified spectrogram in this manner is estimating what the necessary phase information should be. Here, phase propagation theory from the phase vocoder (Flanagan et al. 1966) could be used to ensure phase continuity between adjacent frames of audio. The phase continuity for any single source in the mixture is inherently disrupted by the phase contributions of all the other sources so using some phase propagation techniques may mitigate this. Furthermore, there is an expected relationship between a sinusoidal peak and its neighbouring bins in a magnitude spectrum generated from a windowed fourier transform. In the source separation process, a peak may be recovered without its neighbouring bins and as such the peak lobe is not correctly formed. Further processing could be applied to ensure that all peaks in the magnitude spectrum have suitable neighbouring bins with suitable phases.

CHAPTER 5: MUSIC STRUCTURE SEGMENTATION USING THE AZIMUGRAM IN CONJUNCTION WITH PRINCIPAL COMPONENT ANALYSIS

This chapter presents the third contribution of this dissertation. Here we show how the *azimugram*, a byproduct of the ADRess algorithm, can be used in conjunction with unsupervised machine learning to perform music structure segmentation. It was originally published at the Audio Engineering Society Convention in 2007 (AES 123) and is presented here in its entirety. The paper included co-authors Mikel Gainza, my research colleague who offered advice on formatting and presentation, and Eugene Coyle who acted as my PhD supervisor at the time.

5.1 - ABSTRACT

A novel method to segment stereo music recordings into formal musical structures such as verses and choruses is presented. The method performs dimensional reduction on a time-azimuth representation of audio which results in a set of time activation sequences, each of which corresponds to a repeating structural segment. This is based on the assumption that each segment type such as verse or chorus has a unique energy distribution across the stereo field. It can be shown that these unique energy distributions along with their time activation sequences are the latent principal components of the time-azimuth representation. It can be shown that each time activation sequence represents a structural segment such as a verse or chorus.

5.2 - BACKGROUND

Music information retrieval is concerned with the automatic extraction of multi-level features from audio for the purposes of classification, comparison and segmentation. In particular, musical segmentation algorithms attempt to segment the audio timeline into perceptually salient events, such as the onset of a particular instrument within the piece, or a key, rhythm or tempo change for example. In (Foote, 2000), Foote utilises an audio similarity matrix in order to find the boundaries between different consecutive self-similar segments. Other methods utilise Hidden Markov Models to segment the audio by clustering sequences of timbre states obtained from a dimensionally reduced constant Q representation of the audio (Levy et al. 2006). Goto presents a method which detects the chorus of a song by using a chromagram

representation (Goto, 2003). The method aims to find the chroma vector which repeats most often in the song. In (Logan et al. 2000), the similar segments are detected by using MFCC features from overlapped audio frames. Perhaps one of the most useful forms of segmentation would allow the identification of the formal structural units of a musical piece, such as verses, choruses and bridges for example. Segmentation in this form would have applications in audio thumbnailing as well as fast audio browsing. Significantly fewer algorithms exist for this level of segmentation although (Levy et al. 2006) and (Goto, 2003) do approach this.

5.3 - METHOD

In this paper, a novel approach to structural segmentation is proposed, using the "azimugram" as the mid-level feature representation from which segmentation is derived. The azimugram is a time-azimuth representation of stereo audio which effectively shows the distribution of energy across the stereo field with respect to time. In this highly condensed domain, source location and intensity are clearly identifiable. Common music composition and production techniques often use additional or reduced instrumentation to herald a section transition in a song. This would suggest that source location and intensity will be highly correlated in similar sections within a given song. The distinct advantage of using the azimugram is the fact that it is invariant to both key changes and melodic variation within similar sections.

Dimensional reduction in the form of PCA (principal component analysis) followed by ICA (independent component analysis) (Hyvarinen et al. 2001) is then applied to the azimugram. This combination of PCA followed by ICA is commonly referred to as ISA (independent subspace analysis). ISA has traditionally been used in source separation problems (Casey et al. 2000) and (Fitzgerald et al. 2002) but we show here that the technique has uses in segmentation also. Performing ISA on the azimugram results in a set of J independent basis function pairs where J is an estimation of the number of unique structural components present in the song, typically $J < 5$. Each of the J basis function pairs consists of one azimuth basis function and one time activation function of dimension $r \times 1$ and $t \times 1$ respectively, where $r \times t$ is the dimension of the azimugram. Taking the first pair as an example; the azimuth basis function corresponds to the most recurring energy distribution profile over time. The corresponding time activation function shows the activation sequence of this azimuth basis function. Each successive pair of basis functions will correspond to a unique energy distribution and time activation sequence. This will be illustrated in section 5.3.2. Only the time activation functions are retained for further processing. Each time activation function is then smoothed using a low-pass filter. At this stage, each time activation function already exhibits a significant amount of structural information, whereby each one clearly represents a particular structural unit of the song such as a verse or a chorus. A final process is then applied whereby for any time instant, only the single largest value amongst all J time activation functions is assigned a value

of one and all others a value of zero. This effectively enforces orthogonality between the functions which ensures that only one segment is active at any given point in time. Each of the *J* functions is now an independent binary sequence which represents the on/off sequence of a particular structural component of the song such as a verse, chorus, bridge or solo for example.

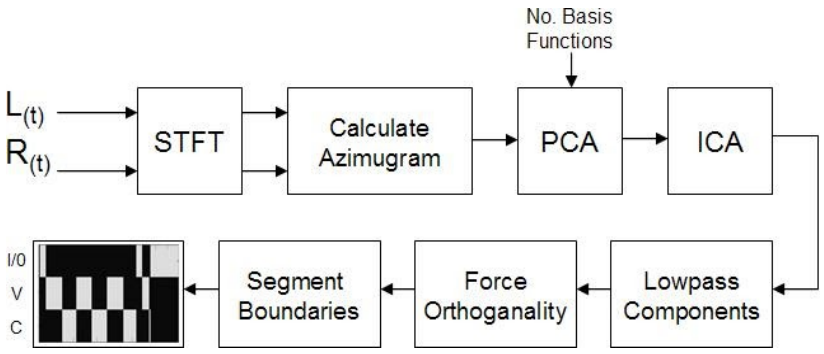

Figure 5.1: Block diagram of the music structure segmentation system.

5.3.1 - The Azimugram

Here, we coin the term *azimugram* to refer to any time-azimuth representation of an audio signal. Such a representation shows the distribution of energy across the stereo field with respect to time. Azimugram representations can be created in various ways depending on the mixing model assumed. Much of the early work concerning azimuth calculation was based on models of binaural perception, whereby the azimugram is calculated by carrying out a cross correlation between the left and right inputs of the system on a multiband basis. The maximum output of the cross correlation functions correspond to the time lag of either the left or right input which can be resolved as an angle of incidence.

An overview of binaural processors can be found in (Stern, 1988). Later work in sound source separation (Barry et al. 2004 b) and (Jourjine et al. 2000), although not explicit, constructed azimugram variants from the short-time Fourier transform of stereo signals. Equations 5.1 to 5.3 below outline a basic technique to calculate an azimugram assuming an intensity stereo mixing model. Firstly, the log ratio of the left and right magnitude spectra is calculated resulting in a matrix of mixing coefficients $A(n,t)$ as in equation 5.1, where $1 \le n \le N$, and N is the analysis frame size. These mixing coefficients are in dB format, whereby positive values refer to components which are dominant in the left channel and negative values refer to components which are dominant in the right channel.

$$A(n,t) = 20Log_{10}\frac{\left|X_1(n,t)\right|}{\left|X_2(n,t)\right|} \qquad (5.1)$$

where, $X_1(n,t)$ and $X_2(n,t)$ are the complex short time Fourier transforms of the left and right channels respectively. Theoretically, $A(n,t)$ will have values in the range of -96 dB to +96 dB for a 16 bit recording where all the positive and negative values correspond to source components dominant in the left and right channel respectively. Following this, a weighted histogram of the mixing coefficients is created on a frame by frame basis. Firstly, the resolution, R, of the histogram is defined, where R specifies how many histogram bins are used to represent each half (left and right) of the histogram.

For example, if R = 32, this will result in 2 x R discrete azimuth locations between far left and far right. Equation 5.2 below, converts the log spaced dB values into linear spaced discrete bin values which are used to populate the histogram created in equation 5.3.

$$\bar{A}(n,t) = R \pm \left(R - \left\lceil \frac{1}{2^{|A(n,t)|/6}} \times R \right\rceil \right) \tag{5.2}$$

where, $2R$ is the resultant histogram resolution and where, $\lceil\ \rceil$ denotes rounding up to the nearest integer. In equation 5.2 above, the term in brackets, preceded by $\pm$, assumes the same sign as the current value of $A(n,t)$. The matrix $\bar{A}(n,t)$ now contains the mixing coefficients in a normalised integer format such that, $1 \le \bar{A}(n,t) \le 2R$. Using equation 5.3, each bin of the histogram, $Az(r,t)$, is then populated by accumulating only the elements, n, of $X_k(n,t)$ where $\bar{A}(n,t) = r$.

$$Az(r,t) = \sum_{k=1}^{2} \sum_{i=1}^{I} \left| X_k\left(B_i, t\right) \right| \tag{5.3}$$

where $B = n \ \forall \ \bar{A}(n,t) = r$

where, $1 \le r \le 2R$, and where k represents the left or right channel indexed by 1 and 2, respectively.

A more accurate way to calculate the azimugram can be found in (Barry et al. 2004 b). This method uses phase information in addition to magnitudes resulting in slightly better localisation for concurrent sources overlapping in time and frequency.

For segmentation purposes, the time resolution of the azimugram must be coarse enough to capture a representative energy distribution for a segment. Typically we use a frame size in excess of 3 seconds with a 50% overlap. Having a finer temporal resolution leads to details of instrument dynamics being exposed which can have adverse effects on the PCA stage used next.

The assumption is that a similar stereo energy distribution can be observed over the course of a single segment, and that the same energy distribution should be apparent whenever that segment is active. In essence, verse 1 is assumed to have a similar stereo field energy distribution to verse 2 for example, and likewise with all other segments.

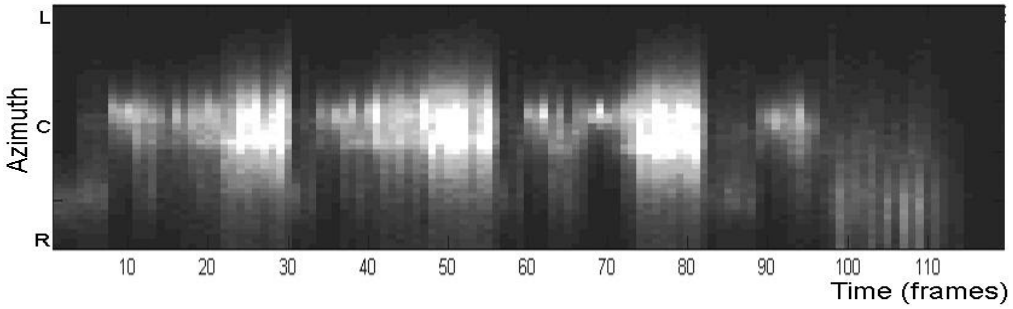

Figure 5.2: Azimugram of Romeo and Juliet – Dire Straits

As stated previously, the distinct advantage of using the azimugram representation is the fact that it is invariant to both key changes and melodic variation within similar sections. Typical values for R are in the region of 20 to 30 points, resulting in an azimuth resolution of $2 \times R$. With this time and azimuth resolution, the azimugram for a 4 minute song would be of dimension 40×160. Such a compact representation facilitates fast segmentation in the following stages.

5.3.2 - Independent Subspace Analysis

The next stage involves performing Independent Subspace Analysis on the azimugram. ISA is a technique used for dimensional reduction which involves performing PCA followed by ICA. The model assumes that the information contained within a data set, in this case the azimugram, can be represented by lower dimensional subspaces, the sum of which approximates the original data set. In the case of the azimugram, each subspace is the result of the product of two latent basis functions of dimension $r \times 1$ and $t \times 1$ respectively, where $r \times t$ is the dimension of the azimugram. Formally stated, it is assumed that the azimugram can be decomposed into a sum of outer products as in equation 5.4.

$$\mathbf{Az} = \sum_{j=1}^{J} Az_j = \sum_{j=1}^{J} r_j t_j^{\mathrm{T}} \qquad (5.4)$$

where T indicates the transpose of the matrix.

In matrix notation, the azimugram **Az**, is represented as the sum of J independent azimugrams, each one corresponding to a particular structural segment of the song.

The basis functions are obtained by carrying out singular value decomposition, commonly known as PCA, on the azimugram. This essentially transforms a high dimensional set of correlated variables into some number of lower dimensional sets of uncorrelated variables which are known as the principal components. The principal components are ranked in order of variance, so the first principal component contains the maximum amount of total variance present in the azimugram and each subsequent principal component represents the maximum remaining variance in the azimugram.

Referring to equation 5.4, the principal components are represented by r_j and t_j. These basis function pairs represent the stereo field energy distributions and the time activations of each distribution respectively. One of the known issues with using PCA is that of choosing how many principal components to use to represent the data. In this application, the number of components, J, is set to be the expected number of recurring structures within the song. Typically, we use 3 principal components, expecting that there will be verses, choruses and other, where other will represent anything which is not a verse or chorus. Of course many other possibilities exist in musical composition, but 3 components should be sufficient to express the general structure of a typical pop/rock song (Covach, 2005).

In order to perform segmentation, only the time activation functions, t_j , are retained. At this stage, the time activation functions are decorrelated but not independent. A limited amount of structure is already apparent within the time activation functions, but there is still activation overlap between the components. Logically, only one structural segment such as a verse or chorus should be active at once, and so theoretically, the basis functions should be mutually exclusive. In order to approach this, ICA is now performed on the time activation functions which results in a set of independent components as opposed to just decorrelated components. Figure 5.3 below shows the first 3 basis function pairs after PCA and ICA.

A known issue with the use of ICA is that the independent components returned can be arbitrarily scaled and/or sign inverted. For this reason, the independent components are normalised and positively oriented before proceeding to the next stage of processing. Following this, a lowpass filter is applied to each of the time activation functions in order to avoid the detection of short segments in the next processing stage. Another issue associated with the use of ICA is that the components could be returned in any order. For segmentation purposes, the components are ordered chronologically, i.e. in the order of time activation. We will refer to these normalised and lowpassed independent components as, $\overline{t}_{j(t)}$.

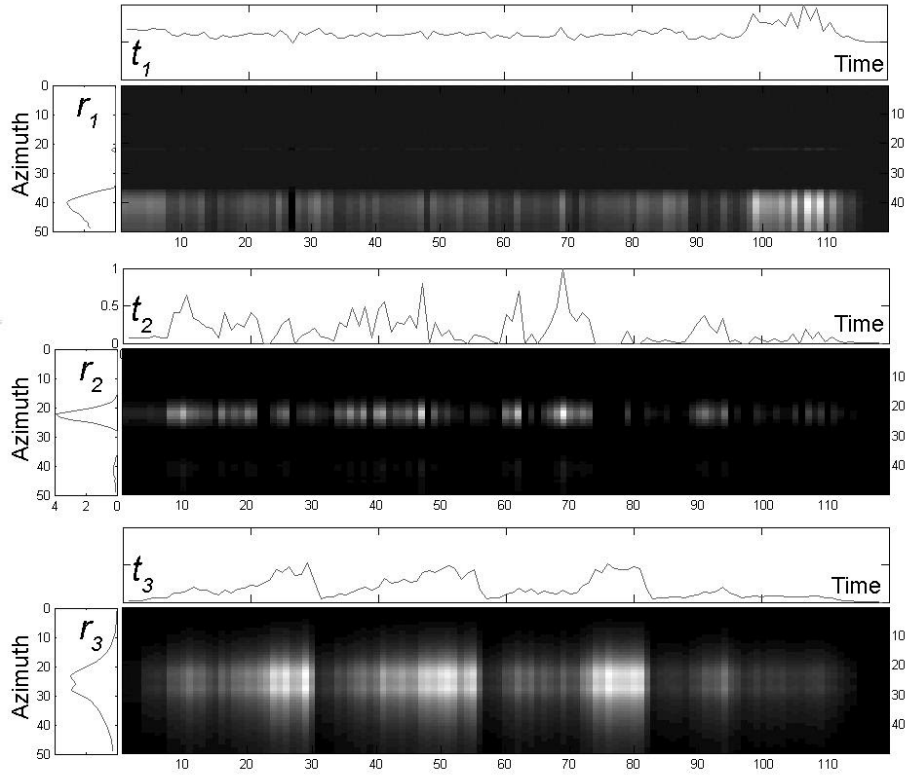

Figure 5.3: The decomposition of the azimugram in figure 5.2 into its first 3 independent subspaces. Here, r and t are the latent azimuth and time activation functions respectively. The independent subspaces are the result of the outer products of each basis function pair obtained using ISA.

5.3.3 - Forcing Orthogonality

At this stage, some structure is apparent from the independent components whereby each component effectively represents the activation of a particular structure such as a verse or chorus but the boundaries between the segments are still unclear. In order to locate the segment boundaries more precisely, the independent components are converted into a set of binary functions by employing an 'all or nothing' scheme whereby for any time instant, the time activation function with the maximum energy is assigned a value of 1 and all others a value of 0 as in equation 5.5.

$$\overline{t}_{j(t)} = \begin{cases} 1 & if \ \ \overline{t}_{j(t)} > \overline{t}_{m(t)} \\ 0 & otherwise \end{cases} \Bigg\} \ j \neq m \qquad (5.5)$$

for $1 \leq j \leq J$, where J is the number of basis functions. This effectively enforces mutual exclusivity. The binary time activation functions now represent the on/off sequence for each structure such as a verse or a chorus. Figure 5.4 illustrates how each stage of processing leads to the resulting structural segmentation.

5.4 - RESULTS

Referring to the example in figure 5.4 above, the frame size was set to approximately 6 seconds with an overlap of 50% resulting in a time resolution of 3 seconds. This essentially means that if a segmentation point is correctly detected within a frame, it will only be accurate to within 3 seconds of the actual segment onset. Analysing Figure 5.4, it can be seen that using PCA alone leaves a significant amount of mutual information in the last 20 frames of the first 2 principal components. Performing ICA in the following stage clearly disambiguates this segment. For this example, the algorithm achieves a high degree of accuracy, correctly identifying the presence of all segmentation points with a maximum error of -6 seconds, corresponding to the early detection of the second chorus.

This is attributed to the fact that the build up into the second chorus is quite prolonged. The instruments are layered more gradually prior to the actual onset of the second chorus.

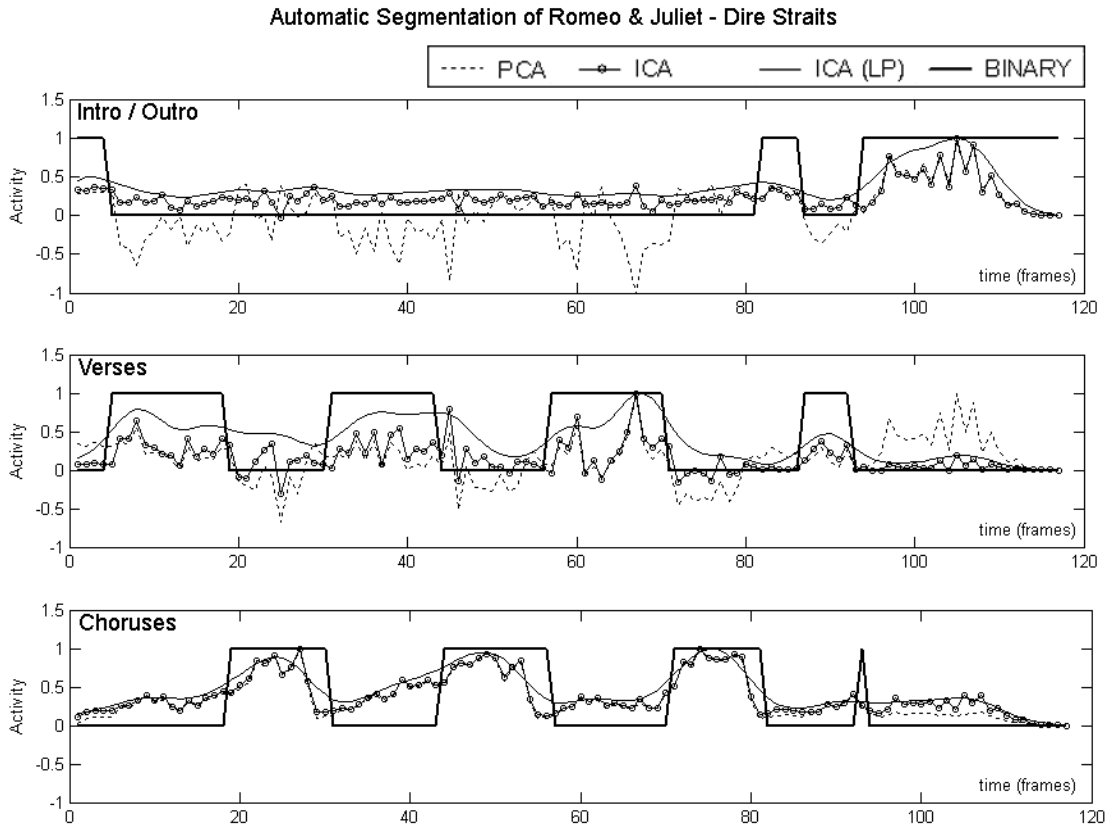

Figure 5.4: First 3 time activation functions after PCA, ICA, lowpassing and binary selection. Note how the functions attain more structure after each stage of processing. Labeling was achieved manually.

This is identifiable from the chorus plot in figure 5.4. Essentially the stereo field distribution at the end of verse 2 is more similar to the distributions observed in the choruses and so has been grouped as such during the PCA stage. The table below shows the automatically generated segment onset times along with the deviation from the manually annotated results. Given that the time resolution used in this example is 3 seconds per frame, the maximum error from the table above, -6 seconds, corresponds to only a single frame error.

Segmentation results for Romeo & Juliet

T	Segment	Actual*	Algorithm*	Deviation*
1	Intro	0:00	0:00	0:00
2	Verse 1	0:22	0:20	-0:02
3	Chorus1	1:05	1:02	-0:03
2	Verse 2	1:39	1:38	-0:01
3	Chorus2	2:22	2:16	-0:06
2	Verse 3	2:56	2:55	-0:01
3	Chorus3	3:39	3:36	-0:03
1	Inst.	4:07	4:09	+0:02
2	Verse 4	4:24	4:24	0:00
3	Error	F.Det.	4:42	N/A
1	Outro	4:46	4:45	-0:01

*time in minutes : seconds

Table 5.1: Comparison of manually annotated segment onset times (Actual) with automatically generated segment onset times (Algorithm). Also indicated is the manually annotated segment name. T indicates the basis function in which the segment was active.

All other segmentation points have been identified within the correct frame with the exception of one false detection at 4:42 which does not correspond to any major structural change. This false detection can be explained by the momentary addition of an ornamental guitar line at that point in the song. The position of this guitar in the stereo field is such that the algorithm incorrectly attributes it to a chorus activation.

The algorithm was also applied to a limited test corpus of popular recordings. The segment onset times for each recording were manually annotated. The automatic segmentation algorithm was then applied to each example and the results were compared. A correct detection was deemed to be within 6 seconds (2 analysis frames) of the manually annotated segment onset. A detection outside this range was considered as an incorrect detection.

Artist	Song	Total Manually Annotated	Correct Detections	Incorrect Detections	Percent Correct
Jimi Hendrix	Castles Made of Sand	8	6	2	75
Busta Rhymes	What's it Gonna Be	7	4	3	57
Whitesnake	Day Tripper	12	8	4	67
Foo Fighters	Everlong	14	10	4	71
AC/DC	Highway to Hell	12	9	3	75
Led Zeppelin	No Quarter	7	5	2	71
Metallica	Nothing Else Matters	9	4	5	44
Fugazi	No Surprise	11	7	4	64
Frank Zappa	Peaches En Regalia	7	4	3	57
Total		87	57	30	65

Table 5.2: Automatically generated segment onset times compared to manually annotated segment onsets.

In this limited test case, the algorithm was able to achieve acceptable segmentation results 65% of the time. Table 5.2 summerises the results obtained.

Although not the focus of this paper, some consideration should be given to the presentation of segmentation data to the user. Figure 5.5 below shows the time alignment of the time domain waveform, the azimugram and a suggested visual representation of structural segmentation. Such a representation gives a user the ability to quickly navigate to important points within the musical piece.

5.5 - CONCLUSIONS

An algorithm capable of achieving automatic structural segmentation on stereo audio signals has been presented.

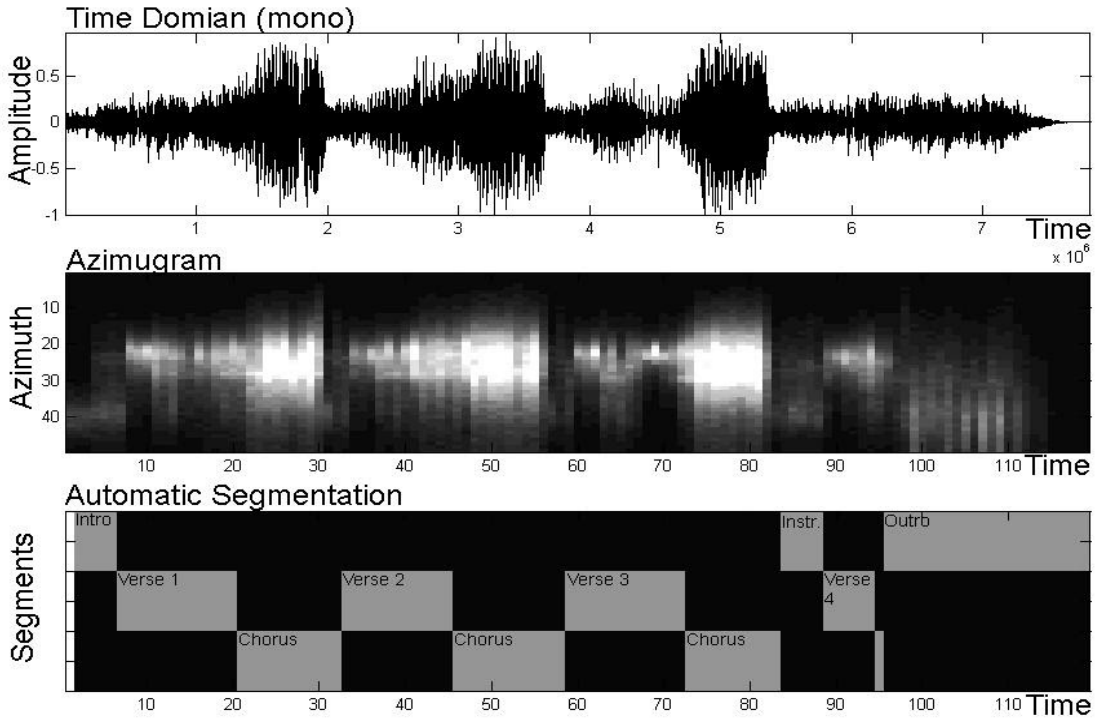

Figure 5.5: Time domain, azimugram and automatic segmentation of Romeo and Juliet from Dire Straits.

The approach is shown to work well on intensity stereo recordings and to a lesser degree on convolutive recordings. The clear advantage of using the azimugram as the mid-level representation, is that it is invariant to key and melodic modulation which is common in music composition. Several problems still exist with the technique however. There is still a difficulty in knowing the exact number of principal components to use in the PCA stage. Added to this, the parameters of the lowpass operation after the ICA stage are still set manually.

5.5.1 - Future Work

Other approaches for matrix decomposition such as locally linear embedding and non-negative matrix factorisation may be used instead of PCA. Although the current formulation is not applicable to mono recordings the same segmentation technique may also be applicable to other mid-level representations such as the chromagram for example. At present, the automatically generated segmentation points are near to the actual segment onsets but as yet are not perfectly aligned with lower level musical events such as bar lines or beats. This will be the topic of further work.

CHAPTER 6: LOCALISATION QUALITY ASSESSMENT IN SOURCE SEPARATION BASED UPMIXING ALGORITHMS

This chapter presents the fourth contribution of this dissertation. Here we show how the ADRess algorithm can be used to generate 5.1 Surround Sound mixes using only stereo content as input. It was originally published in the Audio Engineering Society 35th International Conference in 2009 and is presented here in its entirety. The paper included co-author Gavin Kearney who designed the perceptual test software used and helped conduct the listening tests in his dedicated facility in Trinity College .

ABSTRACT

In this paper we explore the source localisation accuracy and perceived spatial distortion of a source separation-based upmix algorithm for 2 to 5 channel conversion. Unlike traditional upmixing techniques, source separation-based techniques allow individual sources to be separated from the mixture and repositioned independently within the surround sound field. Generally, spectral artefacts and source interference generated during the source separation process are masked when the upmixed sound field is presented in its entirety; however, this can lead to perceived spatial distortion and ambiguous source localisation. Here, we use subjective testing to compare the localisation perceived on a purposely generated discrete presentation and an upmix (2 to 5 channel) of the same source material using a source separation-based upmix algorithm.

6.1 - INTRODUCTION

Surround Sound technology has become commonplace in modern gaming and entertainment applications. Whilst a large proportion of audio content is authored specifically for multichannel reproduction, some pre-existing content is often repurposed for surround sound presentation. Upmixing techniques are typically used to generate several reproduction channels from a limited number of source channels. Traditional approaches often involve ambiance extraction, typically through mid-side processing and channel delay schemes to increase immersion in the resultant sound field.

Although these approaches do provide a greater sense of spatialisation, they do not facilitate localisation of discrete sound sources within the surround sound field. Upmixing techniques based on sound source separation algorithms afford the possibility of repositioning sources discretely within the surround field offering greater upmix flexibility.

This study is not concerned with comparing existing separation algorithms for the purposes of upmixing, rather, the purpose of the experiment proposed here, is to subjectively compare the localisation perceived on a purposely generated 5 channel presentation and an upmix of the same source material using a source separation-based upmix algorithm. Purpose generated multi-track recordings are used to create both a 5 channel mix and a 2 channel mix. Using the source separation-based upmix algorithm, the 2 channel mix is then upmixed to emulate the discrete 5 channel mix. Using subjective testing, it is then possible to directly compare the localisation achievable between the purpose generated 5 channel mix and that of the 2 channel upmix. For the experiments we use a modification of the ADRess algorithm (Barry et al. 2004) as the basis for our upmixing model. The algorithm uses a novel spatial clustering and adaptive filtering technique to identify and separate sources in real time based on their location within the stereo field. The sources can then be remixed and/or re-authored with relative ease.

6.2 - BACKGROUND

6.2.1 - Traditional Upmixing Techniques

The origin of up/down-mixing techniques can be traced back as far as the Quadraphonic era, where four discrete channels of audio were encoded onto two channel vinyl discs (Eargle, 1971). The discs accommodated playback on standard stereophonic record players or four channel playback with dedicated Quadraphonic decoders. Unfortunately, due to competing technologies, increased production costs, and a confused public, the Quadraphonic era ended in a complete commercial failure. However, by the end of its demise, the principles of 'matrix' encoding and decoding on which Quadraphonics was founded had already migrated from the domestic environment to the cinematic world. In 1975, Dolby Systems introduced 'Dolby Stereo' (Hull, 1994), a method of encoding four cinematic audio channels onto the two optical channels found at the side of 35mm cinematic film. The original studio master reproduction channels, *L*, *R*, *C*, and *S* (the left, right, centre and surround channels respectively) are encoded onto the L_T and R_T channels of the optical soundtrack. Decoding of the *S* and *C* channels involves the sum and difference of the two optical L_T and R_T channels, such that phase shifted surround components will cancel each other out in the decoded centre channel, and that the centre channel will be removed from the decoded surround channel. This is achieved by several matrix operations as outlined in (Dressier, 1993). A major consequence of such matrixing is the crosstalk inherent in each channel.

Both the surround and centre channel components in the decoded L_{Front} channel are each only 3dB down from the original L component. This is the same for the R_{Front} channel. Crosstalk in the surround channel is overcome by delaying the surround feed such that localisation precedence is maintained towards the three frontal channels.

Pro-Logic, the consumer version of Dolby Stereo, improves image stability somewhat by including active 'logic steering' circuitry which attempts to steer images towards one speaker. The control circuit looks at the relative levels and phases of the input signals in order to control a group of VCAs which govern the antiphase signals in the output matrix. However, in a 5 speaker setup, the VCAs do not control steering in the Left-Right axis and the Front-Back axis separately. In Pro-Logic II (Dolby, 2004), each axis operates individually through inclusion of a feedback servo control system that adjusts the levels of the VCAs controlling the L_T, R_T, L_T+R_T and L_T-R_T signals such that better channel separation can be achieved.

Such matrix encoding and decoding has received marketplace acceptance as the standard for cinematic upmixing, but we must bear in mind that the majority of stereophonic *music* presentations are not matrix encoded. This leads to distinct differences between how Pro-Logic systems handle cinematic and music program material. Music mode in Pro-Logic II systems includes a high-shelf filter in the surround channels, whereas movie mode does not.

There is also no delay component for the rear channels, which although desirable for coincident arrival wavefronts at the centre listening position (in particular transients), can lead to a perceived reduction in channel separation (Dolby 2004)

It is clear that although matrix systems have significantly developed from their beginnings as humble passive decoders into sophisticated solutions for up-mixing from two-channel material, their application to all types of program material is not fully satisfactory. Furthermore, the fact remains, that in order to obtain optimal performance from any matrix system, the two channel material needs be properly preconditioned (encoded) beforehand (Dolby 2004).

6.2.2 - Source Separation And Upmixing

Sound source separation refers to the task of extracting individual sound sources from some number of mixtures of those sound sources. Unlike matrixing technology, the source material does not have to be pre-encoded for effective upmixing to be achieved. In recent years, advances in dual channel sound source separation technology such as the DUET algorithm (Jourjine et al. 2000) and the ADRess algorithm (Barry et al. 2004) have made it possible to achieve high quality separation of individual sources from stereophonic mixtures. The former is applicable for speech separation in spaced sensor convolutive mixtures whereas the latter is designed for separating or ‘de-mixing’ intensity panned (linear mixed) stereophonic music content.

The primary focus in development and application of (Jourjine et al. 2000) and (Barry et al. 2004) above was purely that of sound source separation. However, prior to (Barry et al. 2004), the application of similar techniques specifically for the purposes of upmixing had been developed in Creative Labs (Avendano et al. 2002) where it was shown that the use of weighted time-frequency masking could be applied effectively in multichannel upmixing. More recently, the same algorithms have been applied to upmixing for Wave Field Synthesis applications (Cobos et al. 2008).

It has been shown in the past that these algorithms are capable of adequate source separation but at the cost of both temporal and spectral artefacts when the sources are reproduced in isolation. Objective comparisons of a number of source separation algorithms are presented in (Vincent et al. 2006) and (Vincent et al. 2007). In general however, such artefacts are perceptually masked when the sound field is reconstructed even after manipulation of individual sources. However, if the content is repurposed for surround presentations, the same artefacts can theoretically manifest themselves through spatial distortion and localisation ambiguity. This can be appreciated if one considers that using the aforementioned separation algorithms; a separated source will often contain time varying interference from overlapping sources within the mix. When the separated sources are then relocated in a multichannel presentation, this interference becomes apparent as channel crosstalk which inherently leads to image shifts in the surround field.

The purpose of this paper is to explore the subjective effects of this image shifting by directly comparing a discrete 5 channel mix and an upmix of the same material.

6.3 - UPMIXING MODEL

For this experiment we use the ADRess algorithm (Barry et al. 2004) with the addition of an azimuth windowing function which was suggested in (Avendano et al. 2002). The ADRess algorithm achieves source separation by taking advantage of destructive phase cancellation in the frequency domain. For each frame, m, of a short-time Fourier representation of the signal, one channel is iteratively gain scaled and subtracted from the other in the complex frequency domain after which the absolute value is taken. The resulting array is of dimension $N \times \beta$, where N is the number of frequency points and β, the azimuth resolution, is the number of equally spaced gain scalars between 0 and 1. The operation reveals local minima, due to phase cancellation across the azimuth plane for each frequency component. Using a simple clustering technique, components belonging to a single source are seen to have their minima in a localised region about some gain scalar which ultimately refers to the intensity ratio between each channel, i.e., the pan position of the source in stereo space. By estimating the magnitude of each of the time-frequency minima and only resynthesising those with a desired intensity ratio, a single source may be reconstructed. The original mixture phase information may be used as was shown in (Barry et al. 2005 c).

The process can be summarised as follows with the iterative gain scaling process achieved using equation 6.1 where $X_j(k,m)$ is a complex frequency domain representation of the m^{th} frame of the j^{th} channel (left or right).

$$Az_1(k,m,i) = \left| X_2(k,m) - g(i)X_1(k,m) \right|$$
$$Az_2(k,m,i) = \left| X_1(k,m) - g(i)X_2(k,m) \right| \quad (6.1)$$

where $1 \leq k \leq N$, N being the Fourier transform length, and where $g(i) = i/\beta$, for all i where, $0 \leq i \leq \beta$, and where i and β are integer values. β refers to the number of gain scalars to be used and ultimately gives rise to the resolution achieved in the azimuth plane. The resulting matrix, $Az_j(k,m,i)$, represents the frequency-azimuth plane for the m^{th} frame of the j^{th} channel. Each of k frequency bins will exhibit a local minimum at some index i. It can be observed that the majority of frequency bins pertaining to a single source should exhibit their minima around a singular value for i. These local minima represent the points at which frequency components experience a reduction in energy due to destructive phase cancellation between the left and right channel.

This energy reduction is directly proportional to the amount of energy which the cancelled source had contributed to the overall mixture and so to invert these minima around a single azimuth point should yield short-time magnitude spectra of the individual sources.

To achieve this inversion, we simply subtract the minimum from the maximum of the function in equation 6.1 for each of k frequency bins as described in equation 6.2.

$$A\bar{z}_1(k,m,i) = \begin{cases} Az_1(k,m)_{\max} - Az_1(k,m)_{\min} & \text{if } A\bar{z}_1(k,m,i) = \min \\ 0 & \text{otherwise} \end{cases} \quad (6.2)$$

where 'min' and 'max' refers to the global minimum and maximum of the k^{th} frequency-azimuth function. Note that the inverted frequency-azimuth plane for channel 2 is created in an identical fashion. Now, the instantaneous magnitude spectrum of a single source or subspace at pan position *d,* predominant in the j^{th} channel can be approximated as in equation 6.3.

$$Y(k,m) = \sum_{i=d-H/2}^{i=d+H/2} A\bar{z}_j(k,m,i) \times \left(1 - \frac{2|d-i|}{H}\right) \quad (6.3)$$

where d is the azimuth index, i.e. the pan position of the source for separation and H is the azimuth subspace width which is simply a neighbourhood around the azimuth index.

The second term in equation 6.3 simply creates a linear weighting function such that components further from the azimuth index are scaled down. This essentially creates a triangular separation window along the azimuth axis. As we will see, the properties of this window will allow adjacent azimuth subspaces to be overlapped in such a way as to allow the extraction of, in this case, 5 discrete subspaces for surround presentation.

Y(k,m) is now an $N \times 1$ array containing the short-time magnitude spectrum of a single source or azimuth subspace. For a detailed description of the ADRess algorithm, refer to (Barry et al. 2004).

6.4 - OBJECTIVE TESTING

Although the algorithms described here and in (Hyvarinen et al. 2001) and (Fitzgerald et al. 2002) are capable of perceptually acceptable separations, a certain degree of signal interference from other sources in the mixture is inevitable in each separation. This section describes the theoretical errors which are known to occur in such algorithms. The material objectively evaluated here is the same as that used for subjective testing in section 6.5. In the case of the algorithm described above and used in this experiment, increasing the value of H will result in capturing more of the desired source for resynthesis but will also lead to a lower signal to interference ratio due to time-frequency (TF) overlap between sources. Theoretically, if the sources do not exhibit TF overlap, near perfect recovery of all sources is possible. However, where western tonal music is concerned, a significant amount of overlap can be assumed. Given that equations 6.1 and 6.2 use both phase and magnitude information to estimate the location of each TF point, the inherent TF overlap between sources causes the local minima to spread out from the true source locations. This is referred to as frequency azimuth smearing in (Barry et al. 2004). This can be observed in Figure 6.1, where the inverted frequency-azimuth plane (N=4096, β=100) for a single frame of the stereo audio is shown.

The audio used here is described in greater detail in section 6.5.1 The audio frame contains 5 sources (guitar, bass, drums, vocals and piano) distributed equally across the stereo field. Referring to Figure 6.1, each frequency component has been resolved to a location within the stereo field. Components naturally cluster close to the theoretical source locations but it can be seen that some components are incorrectly localised and so wider subspace widths (H) would be required to faithfully approximate sources at the cost of unwanted interference.

This ultimately means that the source estimates, $\hat{S}_j(t)$, are not equal to the true sources $S_j(t)$ but the sum of the source estimates should be approximately equal to the sum of the true sources as in equation 6.4. This is a known shortcoming of such separation algorithms. Nevertheless, in the case where the stereo presentation is reconstructed, even with individual source manipulation, the artifacts are generally not discernable (Avendano et al. 2003) but the same artefacts could theoretically lead to noticeable localisation ambiguity when reproduced for surround presentation. Section 6.4.2 explores this issue further.

$$\widehat{S}_j(t) \neq S_j(t) \qquad \text{but}\ldots \qquad \sum_{j=1}^{J} \widehat{S}_j(t) \approx \sum_{j=1}^{J} S_j(t) \qquad (6.4)$$

6.4.1 - Reconstruction Errors

The frequency-azimuth smearing illustrated in Figure 6.1 essentially leads to reconstruction errors in each of the individual source estimates.

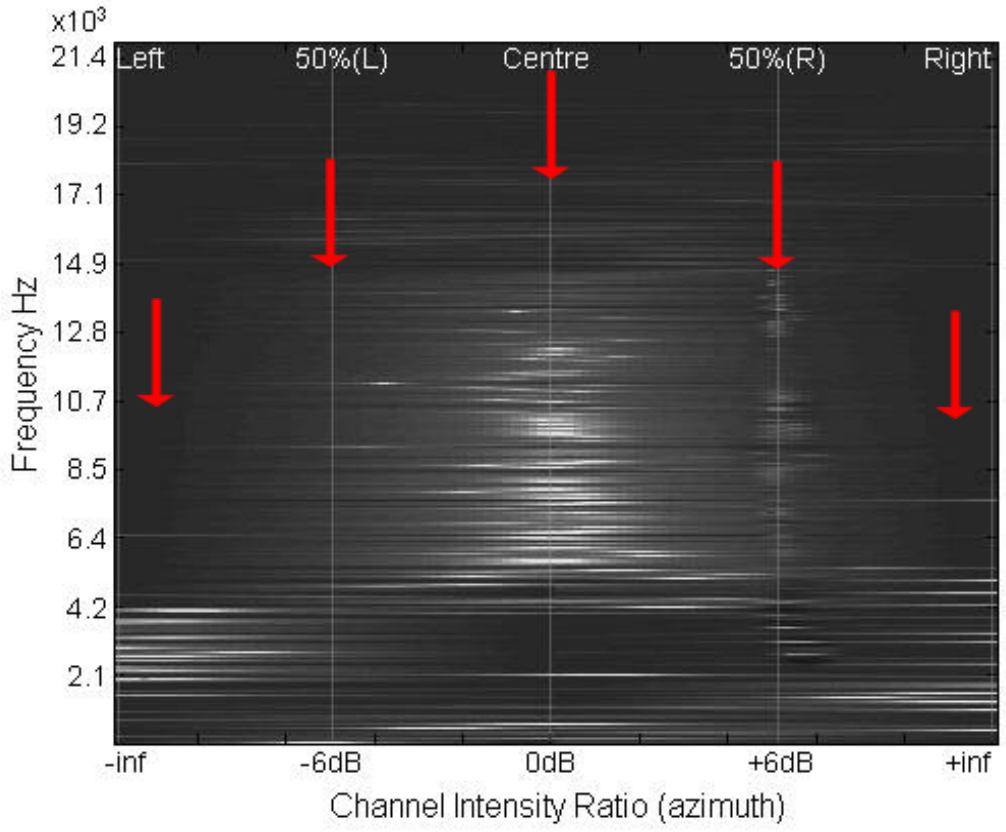

Figure 6.1: Inverted frequency-azimuth plane for a single audio frame as described by equation 6.3. Five sources are clearly present, distributed equally from far left (-inf) to far right (+inf) as indicated by the red arrows. Note the smearing of frequency components across the azimuth plane.

This reconstruction error will depend ultimately on the number of instantaneously active sources and their relative TF overlap. In (Vincent et al. 2007), a set of objective measurement criteria were presented in order to compare the reconstruction quality of a number of source separation algorithms. The criteria proposed were as follows:

- **ISR – Image to Spatial distortion Ratio (dB)**

  This measurement assesses the algorithms ability to estimate the individual source contributions to each channel in the mixture signal.

- **SIR – Source to Interference Ratio (dB)**

  Here, the presence of unwanted interference from other sources in the mixture is measured as a function of the source estimate itself.

- **SAR – Source to Artifact Ratio (dB)**

  Additional algorithm specific artifacts are also measured as a function of the source estimates.

- **SDR – Signal to Distortion Ratio (dB)**

  This measurement conveniently combines all error measurements described above. Refer to (Vincent et al. 2007) for a detailed description of the derivation of these measures.

In order to have some objective measures to refer to for comparison purposes, the subjective test material used in section 6.5 has been processed using the blind source separation evaluation toolbox (Févotte et al. 2008) which implements the error measurements described above. Figure 6.2 presents the error measurement criteria for each of 5 source estimates separated from the stereo mix. These 5 source estimates will ultimately comprise the 5 channel upmix in section 6.5. Note, the original implementation uses the $10\log_{10}$ power law for error measurement but here we use the $20\log_{10}$ power law given its prevalence in the audio domain.

Referring to Figure 6.2, it can be seen that the vocal has achieved the greatest amount of separation owing to the fact that it is the most prevalent source in the stereo mix. Subsequently, the bass, the lowest source in the stereo mix achieves the poorest SIR.

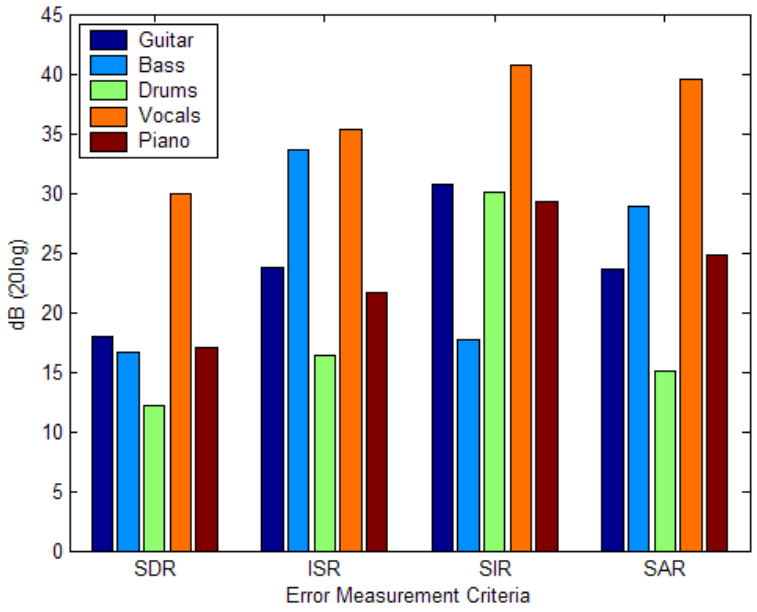

Figure 6.2: SDR, ISR, SIR and SAR for each of the five separated sources from stereo mixture from which the experimental upmix will be generated. Sources positioned from far left to far right as follows: guitar, bass, drums, vocals and piano.

This is a property of almost all separation algorithms, whereby the loudest sources will generally have the greatest influence during clustering stages. Both guitar and piano exhibit similar error values owing to the fact that they exhibit significant TF overlap (between each other) and are of similar amplitude in the stereo mix. In general however, it can be seen that in this example, an average SIR of 30dB can be achieved with a minimum of 17dB in the case of the bass.

6.4.2 - Image Shifting

Given that source separation is generally the task of solving an underdetermined problem, theoretical errors are inevitable as discussed above. As such, we consider the effects of such errors when separation algorithms are used for multichannel upmix.

As described above, interference from nearby sources is the most prevalent problem, whereby an individual source estimate will invariably contain some unwanted components from other sources. Consider the upmix task, where in this case 5 virtual sources from the stereo mixture will be repurposed as 5 discrete sources for a 5 channel presentation. This source interference becomes channel crosstalk which should theoretically result in image shifting within the surround presentation. Subjectively, this should lead to localisation errors.

In order to illustrate how TF overlap causes localisation errors in the separation algorithm we derive the azimugram (time-azimuth representation) of the stereo mix used for upmixing in this experiment. Essentially each column in Figure 6.3 is the transposed column sum of a frame such as that presented in Figure 6.1. Referring to Figure 6.3, note the encircled area, where it can be clearly seen that source overlap has caused the source image to temporarily shift towards the centre. This theoretical error will result in channel crosstalk in any subsequent upmix of the material.

6.5 - SUBJECTIVE TESTING

A subjective experiment was designed to compare the localisation accuracy of a 5 channel musical presentation created from an upmix using ADRess against a discrete 5 channel presentation. The aim of this test was to quantify the extent of localisation shifts due to the source interference in the upmixing algorithm. The test was performed in accordance with the ITU BS.1284-1 recommendations for listening tests (ITU, 2002) and conducted on a standard ITU 5-channel layout.

Bass management (where low-frequency content from the main surround channels is routed to a subwoofer) was omitted from this experiment on the grounds that it may bias localisation of lower frequency range sources. In the context of this experiment, we would expect SIR and ISR to be the most useful indicators of spatial distortion in the 5 channel upmix of the source material because they are each a proxy measure for how much the sources have overlapped in the time-frequency domain.

6.5.1 - Material Preparation and Stereo Mix

For the tests, a dedicated 2 channel stereophonic recording of a jazz ensemble was created. The recording consisted of 5 discretely recorded sources; Piano, drums, vocals, electric guitar and bass. The recordings are of studio quality and were taken at 96kHz, 16-bit. A stereo mix of the sources was generated such that the 5 sources were distributed equally across the stereo stage giving 5 equal width source subspaces that could be separated to produce the 5 channel upmix. The mixing criteria for the stereo mix is shown in Table 6.1.

The spectral contribution and relative mix intensity of each source can be seen in Figure 6.4. The drums are the most spectrally dense source, whilst the vocals contain the most significant energy in the mix. The bass guitar has the most limited frequency range with prominent spectral components below 300Hz.

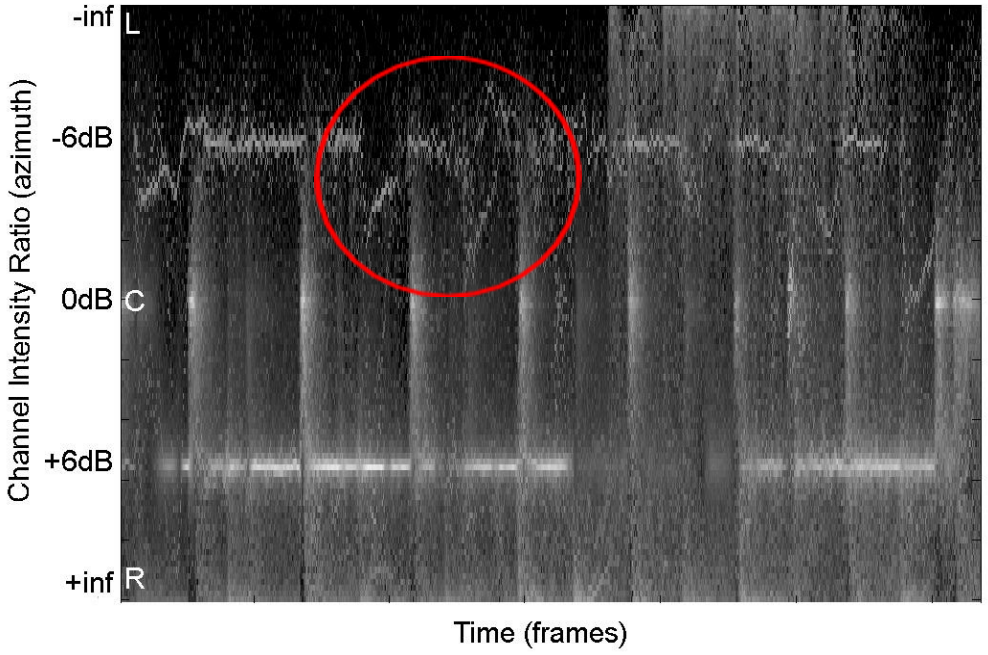

Figure 6.3: The time-azimuth representation of several hundred audio frames. Source activity is clearly visible as is source overlap leading to localisation errors in the source separation algorithm.

Instrument	**Level**	**Pan Position**
Guitar	-5.8 dB	Left (100%)
Bass	-8.7 dB	Left (50%)
Drums	-7.2 dB	Centre
Vocals	0 dB	Right (50%)
Piano	-6.4 dB	Right (100%)

Table 6.1: Mixing parameters for stereo mix. Level measurements are normalised and averaged over 200mS frames where all 5 sources are present simultaneously.

6.5.2 - Upmixing

In any 5 channel upmix, there are two-main methods of placing the audio sources (Avendano et al. 2002). These are 'audience-view' (where the sources are kept at the front of the surround array and the rear speakers are used for lateral spatial enhancement), and 'ensemble view' (where the listener is put in the centre of the musical presentation, surrounded by the musical sources). The first approach is akin to ambience extraction, which is not the focus of this work.

Here we adopt the latter approach, where we attempt to separate 5 equal width, overlapping, azimuth subspaces from the stereo field (see Figure 6.5) so that each source might be uniquely mapped to a single loudspeaker in the 5 channel upmix. The modified ADRess algorithm described in section 6.3 was used for this purpose.

6.5.3 - Experimental Procedure

It was the task of each participant to attempt to identify the direction of the upmixed sources. For the upmix, there are 120 possible permutations by which all 5 sources can be mapped to the loudspeakers. However, we can limit the number of tests such that we are only interested in permutations where we can test localisation of each source uniquely mapped to each loudspeaker. Thus we only need to construct 25 different tests. This can be further reduced if we consider the symmetry of the array, since symmetrically equivalent tests should give identical results.

This results in 15 unique tests with which to describe the localisation accuracy of the upmix. Also, for each upmix, there is then an exact discrete channel mix with which to compare the localisation accuracy, giving a total of 30 localisation tests for each participant. In total, 10 listeners were chosen for the tests, each under 35 years of age, of excellent hearing, and well experienced in musical production. The setup illustrated in Figure 6.6 consists of 5 Genelec 1029A loudspeakers each calibrated to 79 dBC at the centre listening position. A MOTU 896-HD audio interface was used to route the audio to each of the loudspeakers and the test was controlled by the participant via a PC laptop. The listening room is a good monitoring environment with a spatially averaged reverberation time of 0.3 seconds at 1kHz.

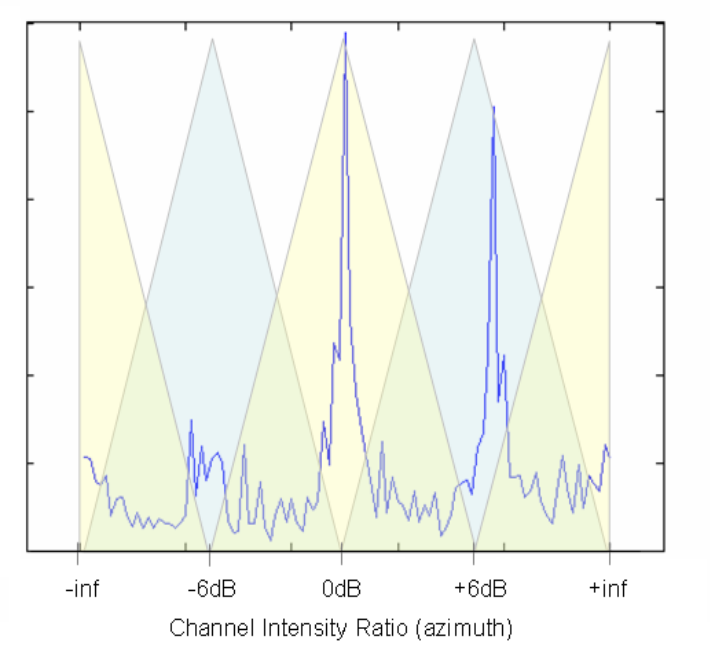

Figure 6.5: Stereo energy histogram illustrating the energy distribution across the stereo field from left (-inf) to right(+inf) within the stereo mix. ADRess is configured to separate 5 equal width overlapped subspaces for upmixing purposes.

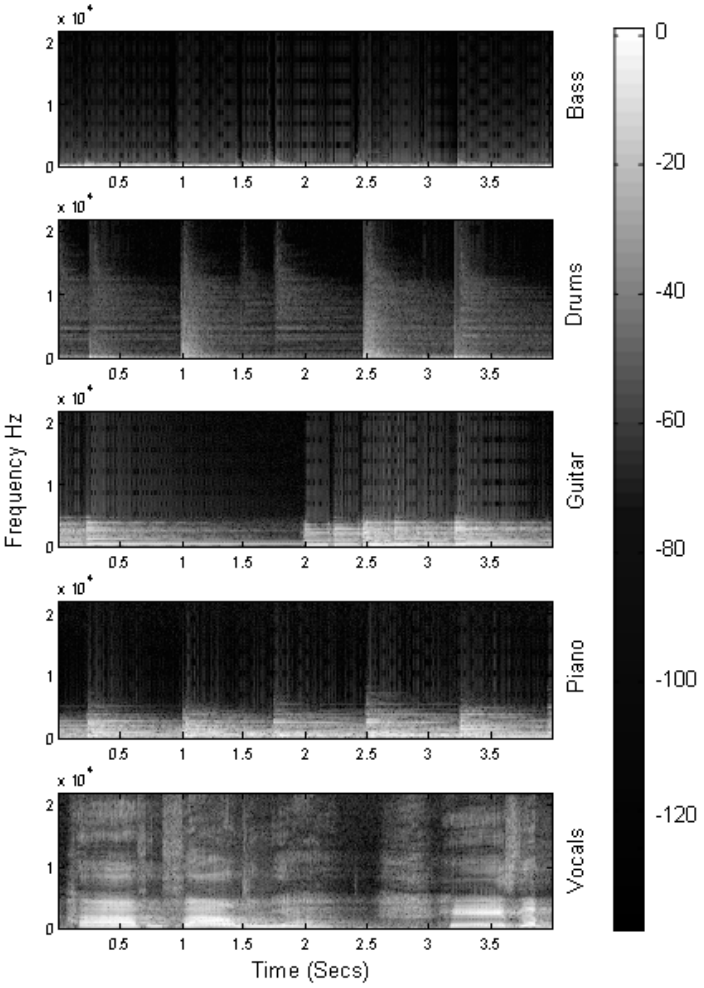

Figure 6.4: Spectrograms of discrete source contributions over 5 seconds of the two channel mix.

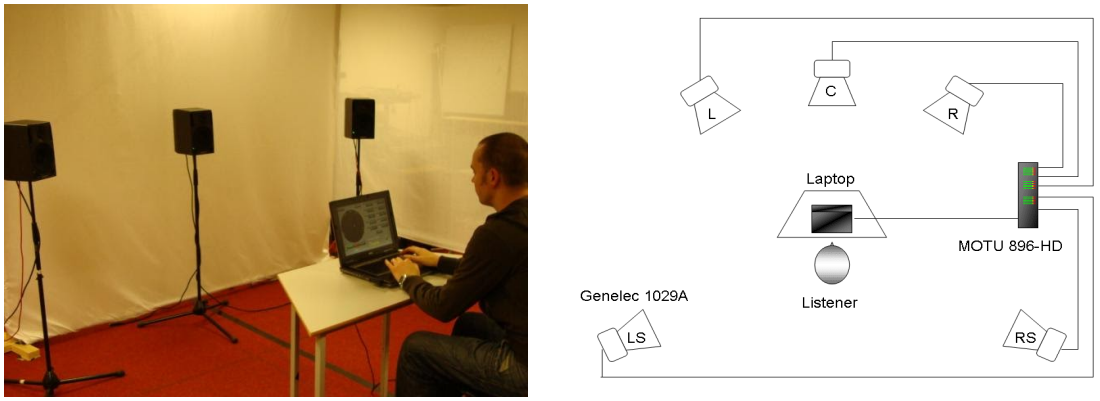

Figure 6.6: Right: Listening Test Configuration. Left: Participant in the listening environment conducting the perceptual experiment with dedicated test software.

6.5.4 - Data Acquisition

A dedicated software pointer, shown in Figure 6.7 was developed to perform the tests. The software gave each participant complete control over the test, allowing them to initiate the audio, stop the presentation or move on to the next presentation. For each test, the software asks the subject to identify the direction of one of the musical sources (shown in large yellow letters). The user can play the test presentation as many times as they desire, before they decide on the direction of localisation using the software pointer. The pointing tool consists of a circle displaying the ITU 5 channel layout with a moveable blue ball for choosing the source orientation. Given the diameter of the ball, there is a 1º margin of error in the test software and the loudspeaker markers are +/-3 º wide. The sequence in which each of the 30 samples is played is completely random and different for each participant.

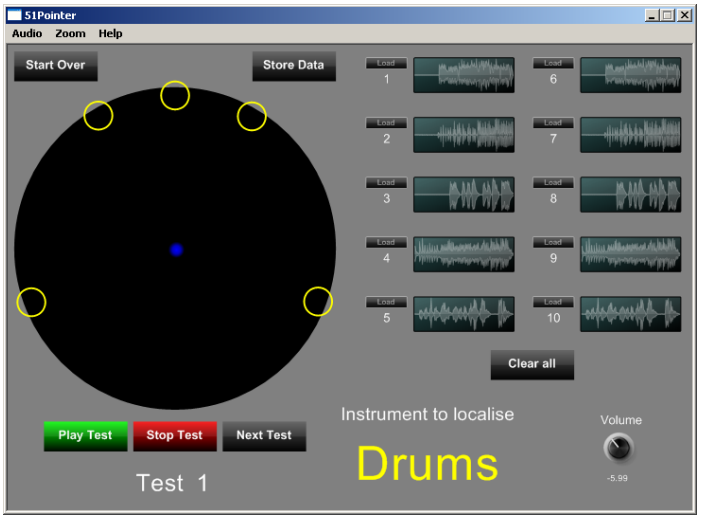

Figure 6.7: Custom software designed for listening test.

6.6 - RESULTS

Observing the results of the subjective testing, it is apparent that the theoretical reconstruction errors discussed in section 6.4.1 have manifested themselves as image shifts within the upmix reproduction. This leads to localisation errors during subjective audition. However, the magnitudes of the errors are dependent on both the instrument and the channel in which it is reproduced. Firstly, we present the data for each reproduction channel (or symmetric pair) as the localisation error from the theoretical source position for each instrument in both the upmix and the discrete mix. Figure 6.8, 6.9 and 6.10 illustrate the perceived localisation error for the center, left/right, and left/right surround channels respectively. Both the discrete 5 channel mix and upmix errors are presented for comparison purposes. Note that 0 degrees refers to the normalised on axis angle for each reproduction channel.

6.6.1 - Center Channel Localisation

Referring to Figure 6.8, it is apparent that the center channel localisation achievable within the upmix is largely similar to that of the discrete mix. Here, the mean localisation error is less than 5 degrees for drums, guitar, piano and vocals. The exception in both discrete and upmix presentations is the bass instrument, where a mean localisation error of 41 degrees and 25 degrees is apparent for the discrete mix and upmix respectively. In general, poor localisation of low frequency content is expected (Theile et al. 1980). Note also that there is an image shift away from the discrete presentation toward the theoretical location.

As a consideration, the SIR for the bass is poorest as indicated in Figure 6.2. This suggests that a substantial number of spectral components from the bass have 'leaked' into other separations. This of course translates to channel crosstalk in the upmix. Thus we postulate that in this case, the crosstalk has affected the perceived localisation of bass within the upmix to positive effect. The complex channel interactions could just as easily result in the opposite effect, shifting the source away from the intended location.

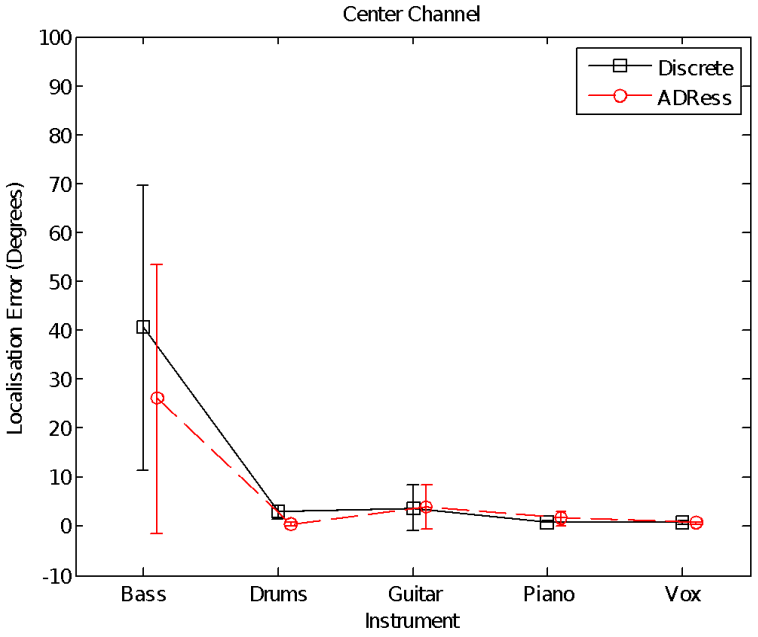

Figure 6.8: Perceived localisation deviations for discrete and upmixed sources positioned in the center channel with theoretical position 0 degrees. (95% Confidence Interval')

6.6.2 - Left and Right Channel Localisation

Referring to Figure 6.9, for left and right channels a noticeable image shift is apparent between the discrete mix and the upmix.

In this case, localisation achievable is clearly poorer for the upmix but the error remains below 10 degrees for drums, guitar, piano and vocals. The bass, as expected, achieves poorest localisation in both cases but a similar situation has occurred whereby the upmix image has been shifted toward the theoretical source location. This has been discussed in the previous section. Note that the vocal has achieved the best localisation. This can be attributed to the fact that it was the loudest source in the stereo mix and achieved the greatest SIR (Figure 6.2) which inherently means that it will generate the least amount of crosstalk in the upmix leading to greater image stability.

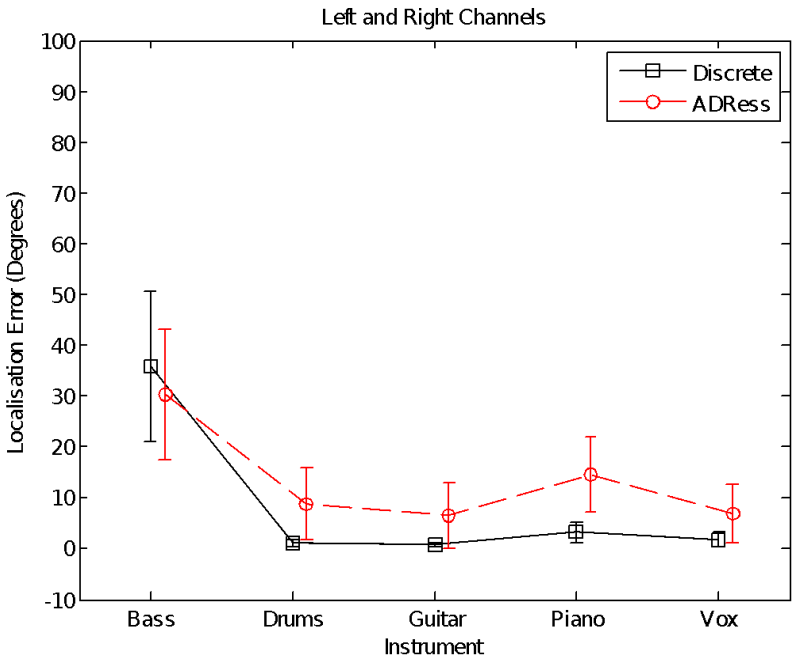

Figure 6.9: Perceived localisation deviations for discrete and upmixed sources positioned in the left or right channels with theoretical positions 30 degrees. (95% Confidence Interval')

6.6.3 - Left and Right Surround Channel Localisation

In general, auditory events presented laterally to a listener are subject to the greatest localisation blur. Blauert (Blauert et al. 1996) shows that sources presented to the sides of a listener undergo, on average, a localisation blur of +/-10 degrees. Both the discrete and upmix presentations illustrate this trait. However, the upmix performs considerably poorer than the discrete mix for rear channels although the trend for each is similar.

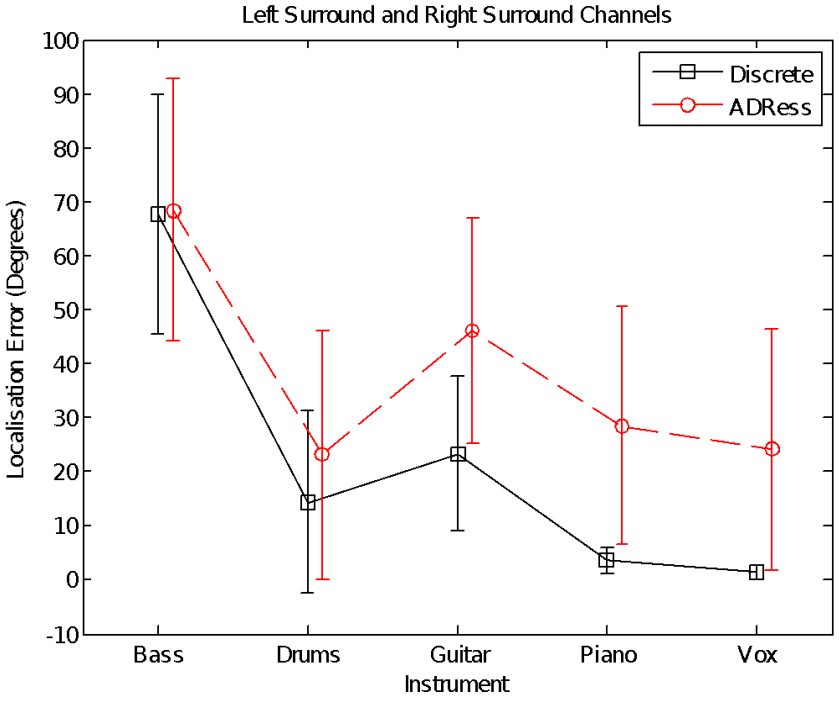

Figure 6.10: Perceived localisation deviations for discrete and upmixed sources positioned in the rear channels with theoretical positions 110 degrees. (95% Confidence Interval')

Note that on average, the upmixed images have shifted 40 degrees from the theoretical positions; however, the shift from the subjective discrete source locations is significantly less, in the region of 25 degrees on average.

Given that the experiment is conducted in a real listening room as opposed to an anechoic chamber, the room acoustics impose constraints on the experiment. We therefore consider the discrete localisation results to be the ground truths as opposed to the theoretical source positions. With this in mind, Figure 6.11 presents the mean image shift of the upmixed source locations as a function of the discrete source locations.

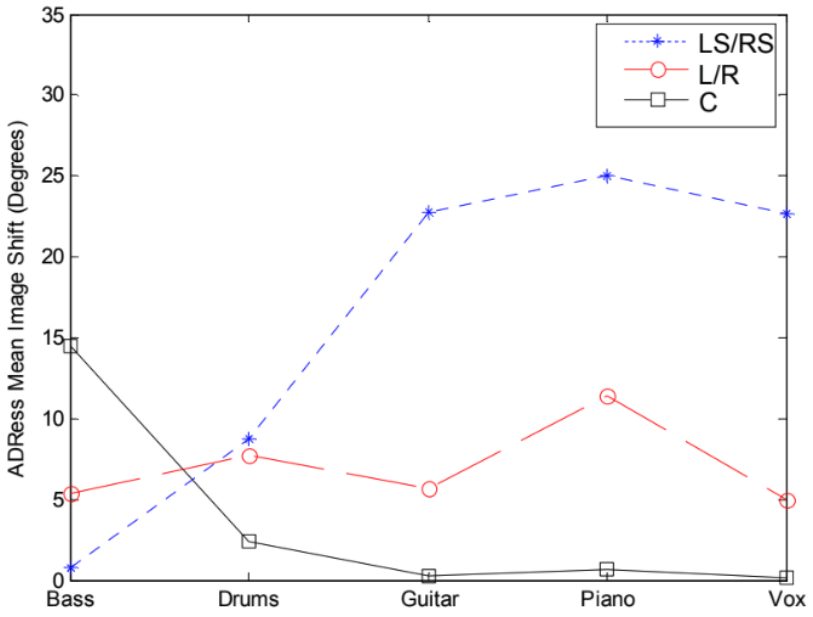

Figure 6.11: The mean image shift observed within the upmix material. (95% Confidence Interval')

6.6.4 - Discussion

In general, the vocal has been localised most accurately in the upmixes with minimum image shifts in the frontal channels. Although the image shift from ground truth is considerable in the surround channels, it remains closer to the theoretical source position than other sources (Figure 6.10).

Subsequently, the vocal also achieves the highest SIR (Figure 6.2) of all sources which implies that it will exhibit less crosstalk upon upmixing. This can be attributed to the fact that the source is almost 6dB louder than any other source in the mix which is advantageous for source separation. Referring to Figure 6.2, the drums achieve the poorest SIR but localisation accuracy remains strong in subjective testing. In general, transients are easier to localise due to the broadband nature of the instruments attack. Secondly, although the drums don't exhibit sustained loudness, they may frequently but briefly become the dominant source in the mixture upon their onset. This aids localisation and would inherently lead to a higher *instantaneous* SIR value. As discussed, bass is difficult to localise in most circumstances. This is evident in both the discrete and upmix presentations. In the case of piano and guitar, they achieve similar localisation accuracy with guitar localisation slightly outperforming that of the piano. This is also supported by the objective measurements where the SIR for guitar is slightly better than that of piano.

In addition to localisation errors, some subjects noted, in rare cases, additional artifacts which were later attributed to upmixed material. Occasionally, some transients were perceived as 'dulled' with respect to the discrete mix although not objectionable. In general, however, many subjects reported that they were often unable to identify which of the two presentations they were listening to in a given test. Finally, it should be noted that in a real-world scenario, the listener has no prior expectation of source locations and so localisation errors are not detrimental to the

effective application of source separation to upmixing, provided that the artifacts known to exist in individual reproduction channels (separations) are masked when the full presentation is recreated.

6.7 - CONCLUSIONS AND FUTURE WORK

In this paper, the source localisation accuracy and perceived spatial distortion of a source separation-based upmix algorithm for 2 to 5 channel conversion was investigated. Subjective and objective testing methodologies were presented in order to assess the localisation accuracy. It was shown that theoretical reconstruction errors associated with the source separation process manifest themselves as image shifts in the upmix presentation and thus led to perceived localisation distortion. However, the localisation error is acceptable in center, left and right channels but significant in the surround channels, yet still below 30 degrees. The tests carried out here are not intended to be comprehensive, but rather, indicative that separation algorithms are suitable for upmix applications, particularly for audience view/ensemble view conversion.

This research focused specifically on assessing the localisation quality of source separation based upmixing presented over a discrete loudspeaker configuration similar to 5.1 surround. Other forms of spatial audio presentation could also be investigated such as wavefield systems and binaural surround presented over headphones. The latter could be assessed for both head-tracked and static presentation.

CHAPTER 7: DRUM SOURCE SEPARATION USING PERCUSSIVE FEATURE DETECTION AND SPECTRAL MODULATION

This chapter presents the fifth and final contribution of this dissertation. Here we present a single channel drum separation algorithm which can be used as a post-process to the ADRess algorithm (Barry et al. 2004) or as a pre-process to drum transcription algorithms such as (Fitzgerald et al. 2002). It was originally published in the IET Irish Signals and Systems Conference in 2005 and is presented here in its entirety. The paper included co-author Derry Fitzgerald who provided the tests on the drum transcription application.

7.1 - ABSTRACT

We present a method for the separation and resynthesis of drum sources from single channel polyphonic mixtures. The frequency domain technique involves identifying the presence of a drum using a novel percussive feature detection function, after which the short-time magnitude spectrum is estimated and scaled according to an estimated time-amplitude function derived from the percussive measure. In addition to producing high quality separation results, the method we describe is also a useful pre-process for drum transcription techniques such as Prior Subspace Analysis in the presence of pitched instruments.

7.2 - INTRODUCTION

In recent years, some focus has shifted from pitched instrument transcription to drum transcription; and likewise in the field of sound source separation, some particular attention has been given to drum separation in the presence of pitched instruments (Helen et al. 2005). Where metadata generation for music archive and retrieval systems is concerned, rhythm analysis is particularly important since broad genre categorization can be ascertained from simplistic aspects of rhythm such as tempo and meter. Automatic drum separation would facilitate more accurate transcription, thus giving access to the finer temporal aspects of rhythm such as polyrhythm and syncopation. Quite apart from this, drum separation and transcription is in itself a useful tool in such applications as computerised music education.

Where the music consists of drums only, some existing algorithms give reasonably accurate results (Fitzgerald et al. 2002), however, in the presence of pitched instruments, the algorithms become less robust and less accurate by way of false beat detection and indeed missing beats altogether (Fitzgerald et al. 2003 b). A drum separation algorithm in this case would be a viable pre-process in order to overcome some of the problems associated with drum transcription in the presence of pitched instruments. Algorithms such as ADRess (Barry et al. 2004 b) and those described in (Avendano et al. 2003) are capable of drum separation in stereo signals if certain constraints are met. In particular, the drums must occupy a unique position within the stereo field. This condition of course is not always met and it is usually the case in popular music that elements of the drum kit share a stereo field position with other instruments. Other algorithms such as (Zils et al. 2002) and (Uhle et al. 2003) have attempted drum separation from single polyphonic mixture signals with varying results. The quality in these cases is usually described as tolerable for the purposes of rhythmic signature analysis. We present a fast and efficient way to decompose a spectrogram using a simple technique which involves percussive feature detection and spectral modulation which results in the extraction of the drum parts from a polyphonic mixture. The algorithm is applicable for the separation of almost any audio features which exhibit rapid broadband fluctuations such as drums in music or plosives, fricatives and transients in speech.

7.2 - METHOD OVERVIEW

Most of the drums used in popular music can be characterised by a rapid broadband rise in energy followed by a fast decay. This is particularly true of the kick and snare drum which could be considered as the most common drums found in modern music. Pitched instruments on the other hand will generally only exhibit energy at integer multiples of some fundamentals which correspond to the notes played in the music. There are of course exceptions in the case of mallet and hammer instruments which may exhibit drum like onsets prior to the stable harmonic regions of the note. With this in mind we develop an onset detector which is not concerned with measuring the rapid rises in energy; but rather an onset detector that measures the broadband nature or *percussivity* of the onset, independent of the actual energy present. In this way drum hits of varying velocity will be detected equally. A percussive temporal profile is derived by analysing each frame of a short-time Fourier transform (STFT) of the signal and assigning a percussive measure to it. The frame is then scaled according to this measure. It should be seen then that regions of the spectrogram with low percussive measures will be scaled down significantly. Upon resynthesis, only the percussive regions remain. Effectively the spectrogram is modulated by an envelope corresponding to the percussion detected within the signal.

Figure 7.1 illustrates the general operation of the algorithm. The magnitude STFT of the signal is taken and the phase information is retained for resynthesis purposes later on.

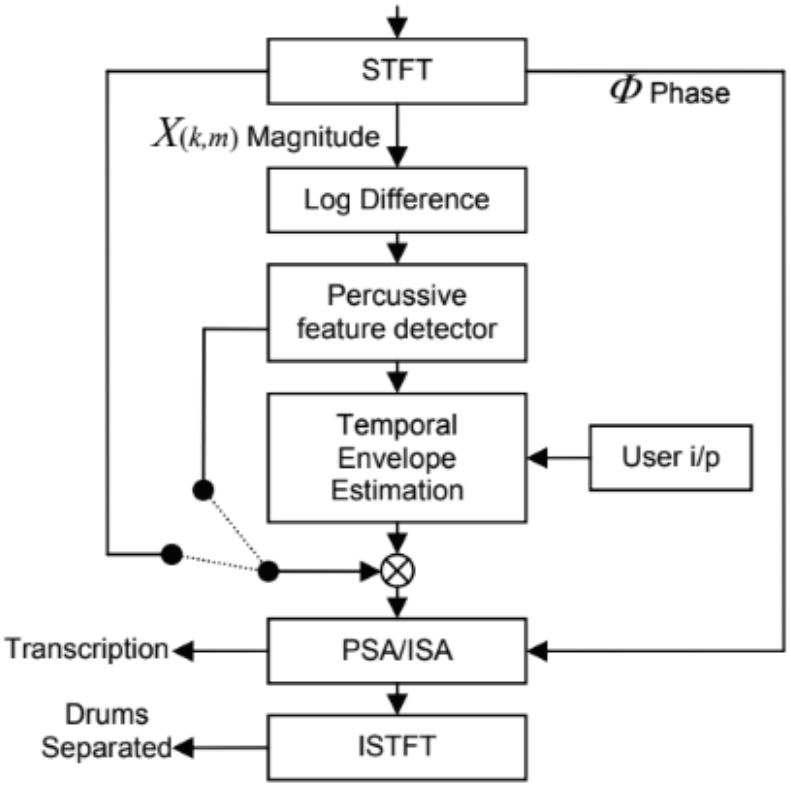

Figure 7.1: System Overview.

The log difference of each frequency component between consecutive frames is then calculated. This measure effectively tells us how rapidly the spectrogram is fluctuating. If the log difference exceeds a user specified threshold, it is deemed to belong to a percussive onset and a counter is incremented. The final value of this counter, once each frequency bin has been analysed, is then taken to be the measure of percussivity of the current frame. Once all frames have been processed, we have a temporal profile which describes the percussion characteristics of the signal. This profile is then used to modulate the spectrogram before resynthesis. Some specific options for resynthesis are discussed in the next section.

7.3 -TEMPORAL ESTIMATION

Firstly we take an STFT of the signal given by:

$$X(k,m) = abs\left[\sum_{n=0}^{N-1} w(n)x(n+mH)e^{-j2\pi nk/N}\right] \qquad (7.1)$$

where $X(k,m)$ is the absolute value of the complex STFT given in equation 7.1 and where m is the time frame index, k is the frequency bin index, H is the hopsize between frames and N is the FFT window size and where $w(n)$ is a suitable window of length N also. Next we take the log difference of the spectrogram with respect to time as in equation 7.2.

$$X'(k,m) = 20\log_{10}\frac{X(k,m-1)}{X(k,m)} \qquad (7.2)$$

for all m and $1 \le k \le K$

In order to detect the presence of a drum we define a percussive measure given in equation 7.3.

$$Pe(m) = \sum_{k=1}^{K}\begin{cases} P(k,m)=1 & if\ X'(k,m) > T \\ P(k,m)=0 & otherwise \end{cases} \qquad (7.3)$$

where, T is a threshold which signifies the rise in energy measured in dB which must be detected within a frequency channel before it is deemed to be a percussive onset. Effectively equation 7.3 acts like a counter; $Pe(m)$ is simply a count of how many bins

are positive going and exceed the threshold. *P(k,m)* contains a '1' if the threshold condition is met and '0' otherwise. Note that the actual energy present in the signal is not significant here; we simply want a measure of how "broadband" or percussive the onset is. The figure below shows the effectiveness of this approach. Standard energy-based onset detectors such as (Masri et al. 1996) will not be able to distinguish between narrowband and broadband onsets. In these systems the level of detection will be intrinsically linked to the energy of the signal at any given time. The detection function we have described is independent of energy and so can deal with low energy onsets as long as they are broadband in nature.

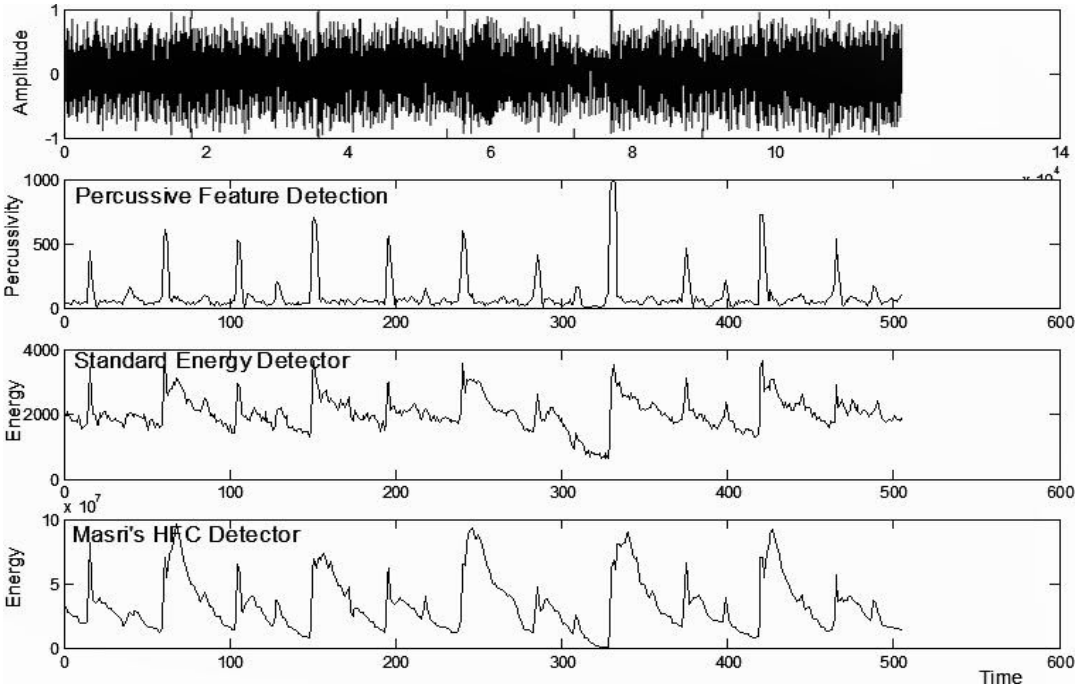

Figure 7.2: The top plot shows the original audio clip. Plot 2 shows our percussive onset detector. The third plot shows the standard energy detector and the bottom plot shows Masri's high frequency weighted detection function (Masri et al. 1996)

Note that the percussive feature detection function we have described even manages to detect the low amplitude hi hat strikes between the kick and snare events.

7.4 - SPECTRAL MODULATION

By weighting each frame by the percussive measure *Pe(m)* , the spectrogram modulates in sympathy with the percussion. This results in the output of the algorithm only becoming active in the presence of a drum sound. There are some options when it comes to resynthesis; the simplest is to simply multiply the original frame by the percussive measure:

$$Y(k,m) = Pe(m)^{\Psi} X(k,m) \qquad (7.4)$$

for all *m* and $1 \le k \le K$

In order to control the decay characteristics of the percussive envelope we simply raise the percussive measure, *Pe(m)*, to the power of Ψ. Larger values of Ψ will lead to faster decay. The parameter is set by the user such that satisfactory results are achieved upon audition. Equation 7.4 results in a time separation of the drum signals but not a frequency separation. Other sources which were present at the same time instant as the drums will also be present but will decay as the drum decays. This method is particularly useful for varying the level of the drums within a mixture signal. For this the separated drum signal is added back to the original signal in some ratio. This process allows for far greater control over the dynamic range of a signal than standard dynamic compression techniques.

The other option for resynthesis which does decouple the drums from the mixture in both the time and frequency domain is as follows:

$$Y(k,m) = Pe(m)^{\Psi} X(k,m) P(k,m) \tag{7.5}$$

By multiplying the frame by the binary mask *P(k,m)*, we are only resynthesising frequency components which were present during the percussive onset. This alters the timbre somewhat but it effectively suppresses non percussive sources in the mixture.

The separated drum signal is then resynthesised using the modulated magnitude spectrum with the original phase information, equation 7.6. It has been shown in (Barry et al. 2005 c) that using the original mixture phase information is more accurate than using a least squared error approximation such as that in (Griffin et al. 1984).

$$y(n+mH) = w(n)\left(\frac{1}{K}\sum_{k=1}^{K} Y(k,m)\cdot e^{j\angle x_{\omega}(k,m)}\right)^{norm} \tag{7.6}$$

The output must be normalised due to the fact that magnitude frames have been scaled according to the percussive measure. *w(n)* is a synthesis windowing function which is required to maintain smooth transitions at the frame boundaries since the process will alter the short-time magnitude spectrum. Since there is both an analysis and synthesis window, it is necessary to use a 75% overlap in order to have a constant sum reconstruction. The algorithm has been applied to many popular recordings and achieves high quality separations in most cases. The figure below shows the separation which has resulted from a typical piece of rock music.

7.5 - RESULTS

The drums are barely distinguishable by visual inspection in the time domain plot on top. However, the percussive feature detector has accurately discriminated between drum events and non drum events. The output of the feature detector is then used to modulate the spectrogram which is inverted to produce the bottom plot which is a time domain reconstruction of the drum events present in the signal.

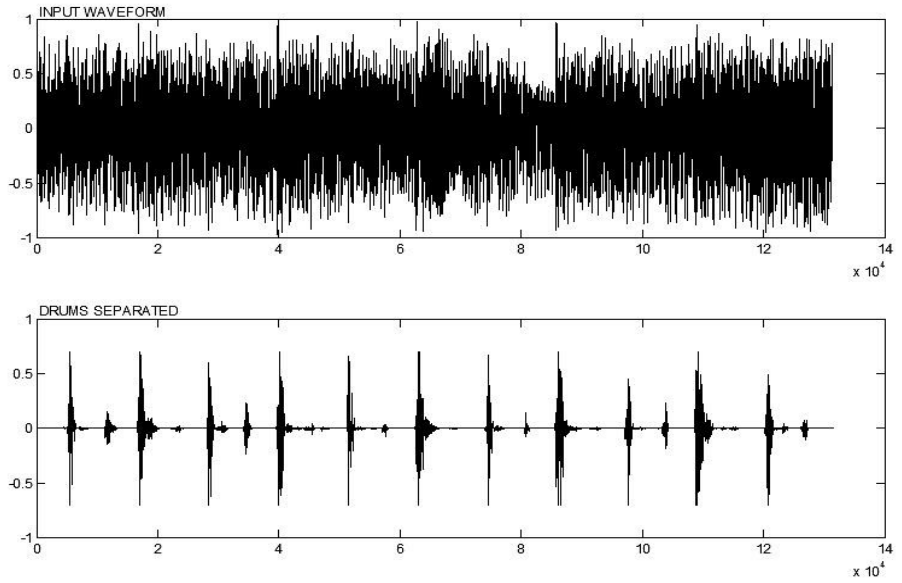

Figure 7.3: The plot shows the original input file and the drum separation which resulted.

To demonstrate the utility of the algorithm as a pre-processing stage before attempting drum transcription, an informal test was carried out on a highly compressed piece of audio which is a "worse case scenario" for drum transcription algorithms. The compression we speak of is dynamic range compression as opposed to bit rate reduction compression. This sort of compression is used to increase the average level of the audio and is applied to many modern recordings in a stage known as 'mastering'.

It effectively reduces peak levels and increases RMS levels dynamically, making it particularly difficult for variance-based transcription techniques such as those in (Fitzgerald et al. 2002) and (Fitzgerald et al. 2003 b) to distinguish the drums at all. The separation algorithm was applied to this recording.

Prior Subspace Analysis (PSA) (Fitzgerald et al. 2003 b), a technique for transcribing drums was then applied to both the unprocessed and separated spectrograms. The results obtained are shown in Tables 7.1 and 7.2. It can be seen that the use of the separation algorithm has substantially increased the performance of the PSA algorithm in transcribing drums in the presence of pitched instruments. The percentages are obtained using the following measure:

$$correct = \frac{total - undetected - incorrect}{total} \cdot 100$$

Type	Total	Missing	Incorrect	%
Snare	5	2	7	-80
Kick	6	1	2	50
Overall	11	3	9	-9

Table 7.1: Drum Transcription obtained using PSA on the unprocessed signal

Type	Total	Missing	Incorrect	%
Snare	5	0	0	100
Kick	6	0	1	83
Overall	11	0	1	91

Table 7.2: Drum Transcription obtained using PSA after the drum separation algorithm

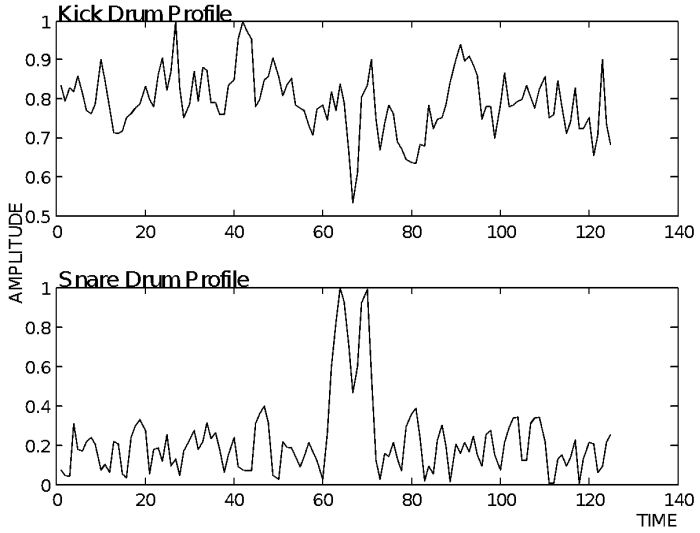

Figure 7.4: ISA was applied directly to the same audio clip shown in figure 7.3.

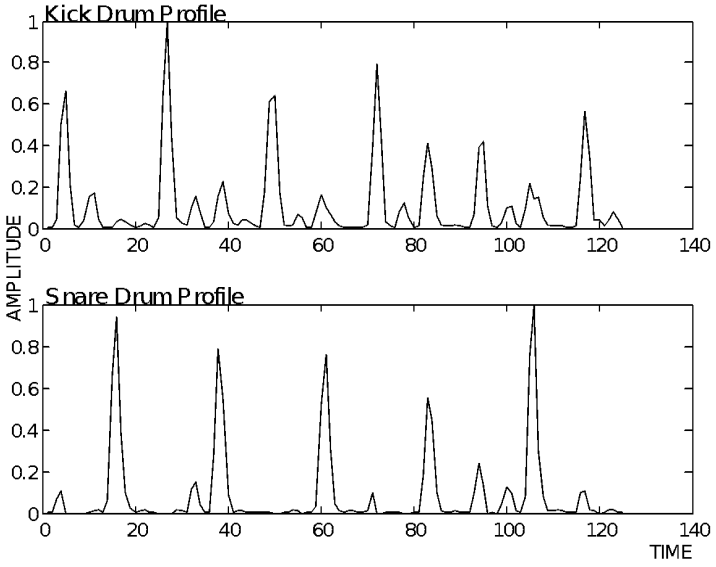

Figure 7.5: ISA after the separation algorithm has been applied

In Table 7.1, the percentage of detection overall is -9% (minus 9%). This was due to the fact that the PSA algorithm made several false positives, i.e. detected events which did not correspond to drum events. 2 out of 5 snares were missed and 1 out of 6 kicks were missed along with several false positives for both.

The results in table 2 clearly show that the PSA algorithm has benefited greatly from the separation technique described in this paper. No events were missed and there was only one false positive in the case of the kick drum.

Independent Subspace Analysis (ISA) techniques (Fitzgerald et al. 2002) also benefit greatly when the separation algorithm presented here is used as a pre-process. The plot in figure 7.4 shows the differences between applying ISA directly to the unprocessed audio, and applying ISA to the separated spectrogram, figure 7.5.

7.6 - CONCLUSIONS

A system capable of separating drum sources from a single polyphonic mixture has been presented. The algorithm is useful in the context of audio processing for music production and education. It has also been illustrated that the use of this algorithm as a pre-processing step for drum transcription algorithms greatly improves the transcription results.

7.6.1 - Future Work

Although the audio quality of the separations is of a high enough standard to be used in the context of transient processing in professional audio applications, the separations played in isolation clearly contain artefacts from other sources active in the non-zeroed time frames. This could be mitigated by trying to estimate the drum spectra more accurately by utilising the fact that the output now contains many

frames, spread across time, which represent a single drum. Given that different melodic events are likely to be playing on each drum hit (kick for example), one could estimate the commonalities and differences across all instances of frames containing a kick drum for example. This could be approached procedurally or using a learning algorithm such as ISA for example (Casey et al. 2000) to extract only the spectral profile of the desired drum.

CHAPTER 8: CONCLUSIONS AND FUTURE WORK

In chapter 2 of this dissertation, a review is presented of the existing sound source separation techniques at the time that the contributions in chapters 3-7 were published. The advantages and disadvantages of each technique are discussed. Surrounding fields of study such as psychoacoustics and cognitive psychology are also explored. Based on this review, a novel algorithm for performing human-assisted sound source separation for music applications, the ADRess algorithm, is presented in Chapter 3. The algorithm is designed specifically to take advantage of the linear stereo mixing model which is sometimes referred to as the intensity panned stereo model. This is the model used by the vast majority of professionally recorded music. Prior to this, most approaches had focused solely on more general cases such as dual or multi microphone source separation or monaural separation. By starting with the desired mixing model we wish to deconstruct, it was possible to tailor an algorithm for that specific purpose. Furthermore, the algorithm is designed to run in real time, a feat that had not yet been achieved at the time of publication. Since its publication in 2004, the ADRess algorithm has had significant impact academically and commercially. The initial two papers and patent have had a total of 177 citations between them and the patent has been cited as prior art by Sony, Samsung, Dolby and NEC on subsequent patents. The algorithm was licensed to Sony in 2006 for use in SingStar on the Sony PlayStation 3 which went on to sell 13m copies. In 2012, the algorithm was licensed to Riffstation, a company I co-founded, which went on to be acquired by Fender and was used by millions of users globally from 2012 to 2018. Also presented are several secondary contributions.

In Chapter 4, an exploration of alternate reconstruction techniques including magnitude-only estimation and sinusoidal modelling were presented.

In Chapter 5, a novel use of the azimugram from ADRess was presented. Here, it is shown that the azimugram can be used, in conjunction with PCA and ICA, to achieve coarse musical structure segmentation. Results are presented for a number of popular recordings.

In Chapter 6, the ADRess algorithm was applied to the task of upmixing. Here, I explored the source localisation accuracy and perceived spatial distortion of an upmix created by the ADRess algorithm. ADRess was configured to carry out a 2 to 5 channel conversion and subjective and objective testing was used to compare the upmix against a dedicated surround mix of the same material. The results show that the typical spectral artefacts that affect single-source separations are not perceivable when all sources are presented over a multichannel playback system such as a 5.1 surround system. However, spatial distortion is perceived but not to an objectionable degree.

In Chapter 7, a novel algorithm for drum source separation is presented. It was originally designed to overcome a shortcoming of the ADRess algorithm; specifically that case where multiple sources are panned to the same location, in which case ADRess cannot separate them. This problem is most apparent in the centre pan position which often contains drums, bass and vocals together. The drum separation algorithm was designed as a post process for ADRess but as shown in Chapter 7, it

was also a very useful preprocess for PSA and ISA-based drum transcription algorithms.

The combined work above was cited 237 times and represented an advance in the field of real-time sound source separation for music applications. This work continues to be extended for other applications today and there is much yet to explore.

8.1 - FUTURE WORK

There are still many avenues to investigate in terms of extending or improving the ADRess algorithm directly. Some of those ideas were suggested at the end of chapters 3 and 4 but are elaborated on here.

8.1.1 Multiresolution ADRess

The analysis frame size for ADRess is chosen to be 4096 samples at 44.1 KHz or approximately 92 ms. This gives an approximate frequency resolution (bin width) of 10.71 Hz. It should be appreciated that this would just about allow the separation of notes spaced 1 semitone apart in lower bass octaves. So in theory, this is quite coarse resolution for low frequencies but more than adequate for high frequencies as the absolute frequency difference between the semitones increases as the fundamental frequency increases. Conversely, this large frame size which accommodates frequency resolution, is detrimental for time resolution. As a result, rapidly changing signals such as transients or other high frequency content will suffer. This is sometimes noticed when drums are separated. Artefacts such as phasiness and transient smearing can often be heard. This happens because the ADRess algorithm seeks to attribute

each frequency component to a dominant source location at any given point in time. It does this by clustering all frequency components within a user defined distance of a specific azimuth. In the case of a transient, it will almost definitely be the dominant source for a very short period of time (significantly shorter than 4096 samples) and should have many frequency components attributed to it by the algorithm but the frame size means that the transient will compete with many other sources which may have dominance at an instant in time either before or after the transient, but within the 4096-sample window. In summary, the time resolution afforded to transients by using a 4096-sample frame size is not sufficient for predictably high-fidelity reconstruction but a smaller window size hinders low frequency reconstruction. The logical solution is to use a multiresolution approach which would split the signal into two or more frequency bands and process each band with a more suitable window size. The processed bands would then be recombined to create the final output. The benefit of this is that the audio fidelity should in theory be perceptually better but the computational requirements will certainly increase.

8.1.2 Inpainting

Inpainting is a concept more often related to image signal processing. It is used to recreate missing or corrupted data in images and it can even be used to synthesise additional content that wasn't present in the original image. Figure 8.1 shows an example of what is possible using machine learning algorithms such as those presented in (Yu et al. 2018) and there are several many more similar algorithms designed for more specific use cases.

Figure 8.1 Example of inpainting results of the method presented in (Yu et al. 2018) on images of natural scene, a face and a texture. Missing regions are shown in white.

Given that audio can easily be represented as a 2-dimensional image in the form of a spectrogram, it is entirely possible to use or at least gain some inspiration from these image-based techniques to fill in missing or corrupted data in audio signals. The inpainting concept would be very applicable to the sorts of artefacts which appear in the ADRess outputs. Prior to the overlap add process in ADRress, it is easily observed that many of the frequency bins will be zero valued. This is due to the central operation of the algorithm whereby some frequency components will be localised near the source of interest and therefore resynthesised, and some will not. Those frequency components which are not localised with the source of interest are set to zero in order to minimise interference from other sources. In theory, any natural source will have some energy at all frequencies (noise at the very least) so estimating those values should contribute to a more natural resynthesis than is currently being achieved. Inpainting could use temporally and spectrally proximate data in the resynthesised spectrogram in order to "guess" at a better approximation for the zero values. This was explored briefly for the ADRess algorithm in (Fitzgerald et al. 2012). Here, NMF was used to achieve inpaintinging. NMF allows for a linear parts-based decomposition of the spectrogram which effectively captures repeating

parts across the whole signal and represents them as a set of global time activation and frequency basis functions. This NMF decomposition was applied to the ADRess outputs in order to achieve inpainting. The idea here is that if certain events recur throughout the signal, each occurrence will have slightly different artefacts due to the interfering sources at that time. By applying NMF, the goal is to generalise the audio events so that parts of the event which were missing at one point in time can be recovered from another point in time where they were not missing through the matrix factorisation process. The results provided minor benefits in some cases but introduced new artefacts in other cases. Although the NMF method may not have been as successful as desired at achieving inpainting, I would strongly encourage exploring similar machine learning methods to achieve the ultimate goal of estimating the data which was not recovered by ADRess.

8.1.3 Peak Lobe Reconstruction

This can be considered as a special case of contextual inpainting. The ADRess algorithm treats each frequency component independently and assumes no mathematical relationship between them. This works surprisingly well, but consider the case of a sinusoidal peak: typically, a sinusoidal peak in the Fourier domain consists of a peak magnitude value in a single bin surrounded by flanking values in neighbouring bins which constitute the lobes of the peak. The lobes also have a specific phase relationship with the peak. If you try to resynthesise a sinusoid from its Fourier representation without its lobes, artefacts will be present. By its nature, the ADRess algorithm does this regularly. It would be possible to include some logic which would attempt to decide if the current frequency component is a sinusoidal

peak, and if so, resynthesise the whole peak including its lobes. This would be worthy of further exploration.

8.1.4 Better Phase Estimation

Chapter 4 explored two alternative phase reconstruction techniques but they are shown to be inferior to the original mixture phases for all its resynthesized sources. This is adequate for a satisfactory result in most cases but it is almost certainly not theoretically accurate. i.e., the mixture phases are not what the individual contributing source phases would have been prior to mixing. Using the nomenclature of the DUET algorithm, if the sources were W-disjoint orthogonal at any point in time, then yes the mixture phase would be a good, if not precise, approximation of the individual source phases, but music is rarely W-disjoint orthogonal (WDO). However, it may be possible to tell which frequency bins belong to WDO sources by analysing the azimuth histogram for each time frame. The azimuth histogram shows a distinctive peak for each source where the L/R channel intensity ratio is identical for many frequency components. Intuitively, we would imagine that it is highly unlikely for many frequency components to share exactly the same channel intensity ratio unless they were related. Further, if those frequency components had been the result of additive energy from multiple sources combining, both the phase and amplitude contributions would almost certainly change the channel intensity ratio for that frequency component. Therefore, we could intuit that only those frequency components with a channel intensity ratio at exactly the source location peak in the azimuth histogram are belonging to W-disjoint orthogonal sources. And therefore, only for those frequency components could we expect the original mixture phases to

be accurate. For all other frequency components within our azimuth window, we should expect the original mixture phases to be a suboptimal estimate of the source phase. For these components, it is worth exploring better phase estimation techniques.

8.1.5 Post Processing

Most of the suggested techniques above aim to modify the operation of the ADRess algorithm internally before resynthesis but there are post processes worthy of exploration. Firstly, it should be reiterated that ADRess separates a source based on its pan position in stereo mix and so in many cases, more than one source might be present in an ADRess separation due to the fact that multiple sources were panned to the same location. For this common case, it should be obvious that monaural source separation techniques could be used as a post process to the ADRess algorithm. Beyond that however, other techniques which avail of multichannel data can also be used as a post process. For example, ADRess could be used to convert a 2-channel stereo mix to 5 individual sources or a stem mix. This multichannel representation could then be post processed by ICA or NMF for example.

8.1.6 Machine Learning

Despite the suggested future work above, I would expect that the future of sound source separation is in machine learning. At the time this research was taking place, machine learning was not nearly as practical as it is today, although unsupervised learning algorithms such as PCA and ICA were gaining traction. It was impractical in terms of published material, public datasets and source code, but also in terms of the development infrastructure supporting it. Now, a TensorFlow model can be trained

and run without the need to install any complex development environment or acquire GPUs for optimised processing. The entire service is conveniently supplied in the cloud including public training data sets in many cases. Many algorithms already in widespread use in the area of image signal processing could be modified to process audio in either the time domain or the time-frequency domain. Particularly in the field of monaural source separation, machine learning techniques such as *variational autoencoders*, *convolutional neural networks* and *recurrent neural networks* can all be used to great avail in the time-frequency domain. DeepMind's Wavenets algorithm has also been applied to sound source separation directly on the time-domain representation of the signal (Lluis et al. 2018). Although these techniques are still only producing comparable results to traditional signal processing techniques, they seem to be advancing faster than the field has ever done in the past. My intuition is that traditional signal processing algorithms like ADRess could be used as a very effective preprocessing step for machine learning techniques to really excel.

REFERENCES

(Abdallah et al. 2003) Abdallah, S.A. and Plumbley, M.D. (2003). An Independent Component Analysis Approach to Automatic Music Transcription. 114th Audio Engineering Society Convention. Amsterdam.

(Allen et al. 1977) Allen, J.B. (1977). Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform. IEEE Transactions on Acoustics, Speech and Signal Processing. 25(3): pp 235--238

(Araki et al. 2007) Araki, S., Sawada, H., Makino, S. (2007). Blind Speech Separation in a Meeting Situation with Maximum SNR Beamformers. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

(Arberet et al. 2006) Arberet, S., Gribonval, R., Bimbot, F. (2006). A Robust Method to Count and Locate Audio Sources in a Stereophonic Linear Instantaneous Mixture. International Conference on Independent Component Analysis and Blind Source Separation (ICA). pp 536–543

(Avendano et al. 2002) Avendano, C., Jot, J. M. (2002). Frequency-domain Techniques for Stereo to Multichannel Upmix. Audio Engineering Society 22nd International Conference on Virtual, Synthetic and Entertainment Audio. Espoo, Finland. 121−130

(Avendano et al. 2003) Avendano, C. (2003). Frequency Domain Source Identification and Manipulation in Stereo Mixes for Enhancement, Suppression and Re-panning Applications. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY. October 19-22 55-58

(Barry et al. 2004 b) Barry, D., Lawlor, R., Coyle, E. (2004). Real-time Sound Source Separation using Azimuth Discrimination and Resynthesis. 117th Audio Engineering Society Convention San Francisco, CA, USA. October 28-31

(Barry et al. 2004) Barry, D., Lawlor, R., Coyle, E. (2004). Sound Source Separation: Azimuth Discrimination and Resynthesis. 7th International Conference on Digital Audio Effects, DAFX 04. Naples, Italy.

(Barry et al. 2005 b) Barry, D., FitzGerald, D., Coyle, E. (2005). Single Channel Source Separation using Short-time Independent Component Analysis. 119th Audio Engineering Society Convention Manhattan, New York, USA.

(Barry et al. 2005 c) Barry, D., Lawlor, R., Coyle, E. (2005). Comparison of Signal Reconstruction Methods for the Azimuth Discrimination and Resynthesis Algorithm. 118th

Audio Engineering Society Convention. Barcelona, Spain. May 28-31

(Barry et al. 2005) Barry, D., Fitzgerald, D., Coyle, E., Lawlor, R. (2005). Drum Source Separation using Percussive Feature Detection and Spectral Modulation. IEE Irish Signals and Systems Conference. Dublin, Ireland. September 1-2

(Barry et al. 2007) Barry, D., Gainza, M., Coyle, E. (2007). Music Structure Segmentation using the Azimugram in conjunction with Principal Component Analysis. 123rd Audio Engineering Society Convention. New York, USA. October 1

(Barry et al. 2008) Barry, D., Dorran, D., Coyle, E. (2008). Time and Pitch Scale Modification: a Real-time Framework and Tutorial. 11th International Conference on Digital Audio Effects (DAFx-08). Espoo, Finland. September 1-4

(Barry et al. 2009) Barry, D., Kearney, G. (2009). Localization Quality Assessment in Source Separation-based Upmixing Algorithms. 35th Audio Engineering Conference. Audio for Games. London. February 1

(Barry, 2019) Barry, D. (2019). Phd Companion Website. https://dan-barry-phd.netlify.com/.

(Bello et al. 2003) Bello J.P., Sandler M. (2003). Phase-based Note onset Detection for Music Signals. IEEE International Conference on Acoustics, Speech, and Signal Processing.

(Bello et al. 2005) Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., Sandler, M. B. (2005). A Tutorial on onset Detection In Music Signals. IEEE Transactions on Speech and Audio Processing. September. 13, pp 1035-1047

(Benetos et al. 2013) Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H. and Klapuri, A. (2013). Automatic Music Transcription: Challenges and Future Directions. Journal of Intelligent Information Systems. doi: 10.1007/s10844-013-0258-3: pp 1-28

(Blauert et al. 1996) Blauert, J. (1996). Spatial Hearing. Revised Edition. MIT Press.

(Blauert et al. 1998) Blauert, J., Grabke, J. W. (1998). Cocktail Party Processors Based on Binaural Models Computational Auditory Scene Analysis. Published by LEA. pp 243-255

(Bodden, 1996) Bodden, M. (1996). Auditory Demonstrations of a Cocktail Party Processor. Acustica. 82: pp 356-357

(Bofill et al. 2001) Bofill, P., Zibulevsky, M. (2001). Underdetermined Blind Source Separation using Sparse Representations. Signal Processing 81. pp 2353–2362

(Bofill et al. 2006) Bofill, P., Monte, E. (2006). Underdetermined Convoluted Source

Reconstruction using LP and SOCP, and a Neural Approximator of the Optimizer. International Conference on Independent Component Analysis and Blind Source Separation (ICA) pp 569–576

(Bregman, 1990) Bregman, A. S. (1990). Auditory Scene Analysis: the Perceptual Organisation of Sound. Cambridge, MA. The MIT Press.

(Brown et al. 1994) G. Brown, M. Cooke. (1994). Computational Auditory Scene Analysis. Computer Speech and Language. 8: pp 297-336

(Casey et al. 2000) Casey, M.A. & Westner, A. (2000). Separation of Mixed Audio Sources by Independent Subspace Analysis. International Computer Music Conference. Berlin, Germany. pp 154-161

(Cherry, 1953) Cherry, E.C. (1953). Some Experiments on the Recognition of Speech, With one and With Two Ears. Journal of the Acoustical Society of America. 25(5): pp 975-979

(Cobos et al. 2008) Cobos, M., Lopez, J. J., Gonzalez , A., Escolano, J. (2008). Stereo to Wave-field Synthesis Music Upmixing: an Objective and Subjective Evaluation. ISCCSP 2008. Malta. March 12-14, pp 1279 -1284

(Covach, 2005) Covach, J. (2005). Form in Rock Music. Engaging Music: Essays in Music Analysis, pp 65-76.

(Dolby, 2004) Dressier, R. (2004). Dolby Surround Pro Logic II Decoder, Principles of Operation. Dolby Laboratories Licensing Corporation.

(Dressier, 1993) Dressier, R. (1993). Dolby Pro Logic Surround Decoder Principles of Operation, Dolby Laboratories Licensing Corporation.

(Duxbury et al. 2003) Duxbury, C., Sandler M. & Davies M. (2003). Temporal Segmentation and Pre-Analysis for Non-linear Time Scaling of Audio. 114th Audio Engineering Society Convention. Amsterdam.

(Eargle, 1969) Eargle, J. M. (1969). Stereo/mono Disc Compatibility: a Survey of the Problems. Journal of Audio Engineering Society. 25355. 17(3): pp 276−281

(Eargle, 1971) Eargle, J. M. (1971). Multichannel Stereo Matrix Systems: an Overview. Journal of the Audio Engineering Society 19(7): pp 552-559

(Ellis, 1992) Ellis, D. (1992). A Perceptual Representation of Audio. MS thesis, Department

of Electrical Engineering and Computer Science, MIT. Cambridge, MA.

(Ellis, 1996) Ellis, D (1996). Predication-driven Computational Auditory Scene Analysis. PhD thesis, MIT Department of Electrical Engineering and Computer Science. Cambridge, MA.

(Ellis, 2003) Ellis, D. (2003). Matlab Implementation of a Sinusoidal Model. http://www.ee.columbia.edu/~dpwe/resources/matlab/sinemodel/.

(Févotte et al. 2008) Févotte, C., Gribonval, R., Vincent, E. (2008). A toolbox for Performance Measurement In (Blind) Source Separation. http://bass-db.gforge.inria.fr, accessed November, 2008. 65-69

(Fitzgerald et al. 2002) FitzGerald, D., Coyle, E., Lawlor. B. (2002). Sub-band Independent Subspace Analysis for Drum Transcription. Digital Audio Effects Conference (DAFX02). Hamburg.

(Fitzgerald et al. 2003 b) FitzGerald, D., Coyle, E., Lawlor. B. (2003). Drum Transcription In the Presence of Pitched Instruments using Prior Subspace Analysis. Irish Signals and Systems Conference. Limerick. July 1-2

(Fitzgerald et al. 2003) FitzGerald, D., Coyle, E. (2003). Independent Subspace Analysis using Locally Linear Embedding. Digital Audio Effects Conference(DAFX03). London, England. 13-17

(Fitzgerald et al. 2005 b) FitzGerald, D., Cranitch, M., Coyle, E. (2005). Shifted Non-negative Matrix Factorisation for Sound Source Separation. IEEE conference on Statistics in Signal Processing. Bordeaux, France. July

(Fitzgerald et al. 2005 c) FitzGerald, D., Cranitch, M., Coyle, E. (2005). Non-negative Tensor Factorisation for Sound Source Separation. Irish Signals and Systems Conference. Dublin. September

(Fitzgerald et al. 2005) FitzGerald, D., Cranitch, M., Coyle, E. (2005). Generalised Prior Subspace Analysis for Polyphonic Pitch Transcription. 8th International Conference on Digital Audio Effects (DAFX05). Madrid, Spain.

(Fitzgerald et al. 2006 b) FitzGerald, D., Cranitch, M., Coyle, E. (2006). Shifted 2d Non-negative Tensor Factorisation. Irish Signals and Systems Conference. Dublin. June 1

(Fitzgerald et al. 2006) FitzGerald, D., Cranitch, M., Coyle, E. (2006). Sound Source Separation using Shifted Non-negative Tensor Factorisation. Proceedings of ICASSP06.

Toulouse France.

(Fitzgerald et al. 2012) Fitzgerald, D., Barry, D. (2012). On Inpainting the Adress Algorithm. 23rd IET Irish Signals and Systems Conference. Maynooth, Ireland. June 28-29

(Fitzgerald, 2004) Fitzgerald, D. (2004). Automatic Drum Transcription and Source Separation. Doctoral Thesis, Dublin Institute of Technology. doi:10.21427/D7002Z

(Flanagan et al. 1966) Flanagan, J.L., Golden R.M. (1966). Phase Vocoder. Bell System Technical Journal. 45: pp 1493-1509

(Fletcher, 1933) Fletcher, H., Munson, W. A. (1933). Loudness, Its Definition, Measurement and Calculation. Bell System Technical Journal. 12: pp 377-430

(Foote, 2000) Foote, J. (2000). Automatic Audio Segmentation using a Measure of Audio Novelty. IEEE International Conference on Multimedia and Expo.

(Goto, 2003) Goto, M. (2003). A Chorus-section Detection Method for Musical Audio Signals. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.

(Gowreesunker et al. 2007) Gowreesunker, B.V., Tewfik, A.H. (2007). Two Improved Sparse Decomposition Methods for Blind Source Separation. International Conference on Independent Component Analysis and Blind Source Separation (ICA).

(Griffin et al. 1984) Griffin D. W., Lim J.S. (1984). Signal Estimation From Modified Short-time Fourier Transform. IEEE Transactions on Acoustics, Speech, and Signal Processing. ASSP-32, no 30774:

(Haas, 1972) Haas, H. (1972). The Influence of a Single Echo on the Audibility of Speech. Journal of the Audio Engineering Society. 20(2): pp 146-159

(Helen et al. 2005) Helen, M. Virtanen, T. (2005). Separation of Drums From Polyphonic Music using Non-negative Matrix Factorization and Support Vector Machine. EUSIPCO 2005, Antalya, Turkey, Sept. 44047

(Howard, 2001) Howard, D.M., Angus J.A.S. (2001). Acoustics and Psychoacoustics. 2nd ed Focal Press, Oxford, Boston.

(Hull, 1994) Hull, J. (1994). Surround Sound Past, Present and Future, Dolby Laboratories Licensing Corporation.

(Hyvarinen et al. 2001) Hyvarinen, A., Karhunen, J., Oja, E. (2001). Independent

Component Analysis. Wiley & Sons.

(Hyvarinen, 2000) Hyvarinen, A. (2000). Survey on Independent Component Analysis. Neural Computing Surveys, http:/ /www.icsi.berkeley.edu/~jagota/NCS. 2: pp 94−128

(ITU, 2002) The ITU Radiocommunication Assembly. (2002). Recommendation ITU-R BS.1284-1 General Methods for the Subjective Assessment of Sound Quality.

(Jeffers et al. 1948) Jeffress, L. A. (1948). A Place theory of Sound Localization. Journ. Comp. Physiol. 41: pp 35-39.

(Jourjine et al. 2000) Jourjine, A., Rickard, S., Yilmaz, O. (2000). Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources From 2 Mixtures. IEEE International Conference on Acoustics, Speech and Signal Processing. June

(Klapuri, 1998) Klapuri, A. (1998). Automatic Transcription of Music. MSc thesis, Tampere University of Technology.

(Kleffner et al. 2007) Kleffner, M.D., and Jones, D.L. (2007). Practical kurtosis−based blind recovery of a speech source in real−world noise. The Journal of the Acoustical Society of America 121.

(Lee et al. 1999) Lee, D.D., Seung, H.S. (1999). Learning the Parts of Objects by Non-negative Matrix Factorization. Nature. 401: pp 788- 791

(Lee et al. 2001) Lee, D.D., Seung, H.S. (2001). Algorithms for Non-negative Matrix Factorization. Advances in Neural Information Processing Systems. 556–62

(Levy et al. 2006) Levy, M., Sandler, M. (2006). New Methods In Structural Segmentation of Musical Audio. EUSIPCO.

(Lluis et al. 2018) Lluıs, F., Pons, J., Serra, X. (2018). End-to-end Music Source Separation: Is It Possible In the Waveform Domain. arXiv:1810.12187 2018. 115: pp 379–391

(Lockwood et al. 2004) Lockwood, M.E., Jones, D.L., Bilger, R.C., Lansing, C.R., O'Brien Jr., W.D.,Wheeler, B.C., Feng, A.S. (2004). Performance of Time and Frequency Domain Binaural Beamformers Based on Recorded Signals From Real Rooms. Journal of the AcousticalSociety of America.

(Logan et al. 2000) Logan, B., Chu, S. (2000). Music Summarization using Key Phrases. IEEE International Conference on Acoustics, Speech, and Signal Processing. 2

(Lunaverus, 2019) Lunaverus. (2019). Available at: Https://www.lunaverus.com/. [Accessed 13 August 2019].

(Mandel et al. 2007) Mandel, M.I., Ellis, D.P.W., Jebara, T. (2007). An EM Algorithm for Localizing Multiple Sound Sources in Reverberant Environments. Advances in Neural Information Processing Systems (NIPS 19).

(Masri et al. 1996) Masri P., Bateman A. (1996). Improved Modelling of Attack Transients In Music Analysis Resynthesis. International Computer Music Conference (ICMC).

(Master, 2003) Master, A. S. (2003). Sound Source Separation of N Sources From Stereo Signals via Fitting to N Models Each Lacking one Source. EE 391 Report. Stanford University.

(McAuley et al. 1986) McAulay, R.J., T.F. Quatieri. (1986). Speech Analysis/synthesis Based on a Sinusoidal Representation. IEEE Transactions on Acoustics, Speech and Signal Processing 34(4): pp 744--754.

(Middlebrooks et al. 1991) Middlebrooks, J.C., Green, D.M. (1991). Sound Localization by Human Listeners. Annual Review of Psychology. 42: pp 135–59

(Mitianoudis at al. 2007) Mitianoudis, N., Stathaki, T. (2007). Underdetermined Source Separation using Mixtures of Warped Laplacians. International Conference on Independent Component Analysis and Blind Source Separation (ICA).

(Moelants et al. 1997) Moelants, D., Rampazzo, C. (1997). A Computer System for the Automatic Detection of Perceptual onsets In a Musical Signal. ICamurri, Antonio (Ed.).

(Mohan et al. 2003) Mohan, S., Kramer, M.L., Wheeler, B.C., Jones, D.L. (2003). Localization of Nonstationary Sources using a Coherence Test. IEEE Workshop on Statistical Signal Processing (SSP).

(O'Grady et al. 2004) O'Grady, P.D., Pearlmutter, B.A. (2004). Soft-lost: Em on a Mixture of Oriented Lines. International Conference on Independent Component Analysis and Blind Source Separation (ICA).

(Paatero et al. 1994) Paatero, P., Tapper, U. (1994). Positive Matrix Factorization: a Non-negative Factor Model With Optimal Utilization of Error Estimates of Data Values. Environmetrics 5: pp 111–126

(Paatero et al. 1997) Paatero, P., (1997). Least Squares formulation of Robust Non-negative

Factor Analysis. Chemometrics and Intelligent Laboratory Systems. Syst.1997,37, 23–35.

(Palmer, 2003) Palmer, S. E. (2003). Visual Perception of Objects. Handbook of Psychology: Experimental psychology. John Wiley and Sons. ISBN 978-0-471-39262-0.

(Plomp et al. 1965) Plomp, R., & Levelt, W. J. M. (1965). Tonal Consonance and Critical Bandwidth. Journal of the Acoustical Society of America. 37: pp 548-560

(Rayleigh, 1907) Rayleigh, L. (1907). On Our Perception of Sound Direction. Phil. 13: pp 214-232

(Rickard et al. 2001) Rickard, S., Balan, R,. Rosca, J. (2001). Real-time Time-frequency Based Blind Source Separation. ICA 2001 Conference San Diego, CA. December

(Rickard et al. 2002) Rickard, S., Yilmaz, O. (2002). On the Approximate W-disjoint Orthogonality of Speech. IEEE International Conference on Acoustics, Speech, and Signal Processing. pp 529–532

(Roman et al. 2001) Roman, N., Wang, D., Brown, G. J. (2001). Speech Segregation Based on Sound Localisation. IJCNN. 4: pp 2861-2866

(Rosenthal et al. 1998) Rosenthal, D. F., Okuno, H. G. (1998). Computational Auditory Scene Analysis. Mahwah NJ. LEA Publishers.

(Roucos et al. 1985) Roucos, S., Wilgus, A. M. (1985). High Quality Time-scale Modification for Speech. IEEE International Conference on Acoustics, Speech, and Signal Processing. pp 493-496

(Ryynanen, 2004) Ryynanen, M. (2004). Probabilistic Modeling of Note Events In the Transcription of Monophonic Melodies. Masters Thesis, Tampere University of Technology, Department of Information Technology. Helsinki, Finland.

(Scheirer, 1998) Scheirer, E. (1998). Tempo and Beat Analysis of Acoustic Musical Signals. Journal of the Acoustic Society of America. January. 103(1): pp 588-601

(Serra, 1997) Serra, X. (1997). Musical Sound Modeling With Sinusoids Plus Noise, From Musical Signal Processing. Swets and Zeitlinger Publishers.

(Slaney et al. 1994) Slaney, M., Naar, D., Lyon, R.F. (1994). Auditory Model Inversion for Sound Separation. IEEE International Conference on Acoustics, Speech, and Signal Processing. Adelaide, Australia. April 19-22

(Smaragdis et al. 2003) Smaragdis, P., Brown, J.C. (2003). Non-negative Matrix Factorization for Polyphonic Music Transcription. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). October. pp 177-180

(Smaragdis, 2013) Smaragdis, P. (2013). Keynote Slides From Waspaa. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).

(Stern, 1988) Stern, R. M., Jr. (1988). An Overview of Models of Binaural Perception. National Research Council CHABA Symposium. Washington, DC.

(theile et al. 1980) Theile, G. (1980). On the Localisation In the Superimposed Soundfield. Dissertation, Technische Universität Berlin.

(Uhle et al. 2003) Uhle C., Dittmar C., Sporer T. (2003). Extraction of Drum Tracks From Polyphonic Music using Independent Subspace Analysis. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003). Nara, Japan. April

(Vincent et al. 2006) Vincent, E., Gribonval, R., Févotte, C. (2006). Performance Measurement In Blind Audio Source Separation. IEEE Transactions on Speech and Audio Processing. 14(4): pp 1462–1469

(Vincent et al. 2007 b) Vincent, E., Sawada, H., Bofill, P., Makino, S., Rosca, J.P. (2007). https://www.irisa.fr/metiss/SASSEC07/

(Vincent et al. 2007) Vincent, E., Sawada, H., Bofill, P., Makino, S., Rosca, J.P. (2007). First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results. International Conference on Independent Component Analysis and Signal Separation.

(Vincent, 2007 b) Vincent, E. (2007). Complex Nonconvex Lp Norm Minimization for Underdetermined Source Separation. International Conference on Independent Component Analysis and Blind Source Separation (ICA).

(Virtanen et al. 2002) Virtanen, T., Klapuri, A. (2002). Separation of Harmonic Sounds using Linear Models for the Overtone Series. IEEE International Conference on Acoustics, Speech, and Signal Processing.

(Xiao et al. 2005 b) Xiao, M., Xie, S., Fu, Y. (2005). A Statistically Sparse Decomposition Principle for Underdetermined Blind Source Separation. International Symp on Intelligent SignalProcessing and Communication Systems (ISPACS).

(Xiao et al. 2005) Xiao, M., Xie, S., Fu, Y. (2005). A Novel Approach for Underdetermined

Blind Source Separation In the Frequency Domain. International Symposium on Neural Networks (ISNN).

(Yu et al. 2018) Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T. S. (2018). Generative Image Inpainting With Contextual Attention. arXivpreprint arXiv:1801.07892.